Received: 2 October 2021 | Revised: 21 April 2022 | Accepted: 5 May 2022

DOI: 10.1111/exsy.13045

ORIGINAL ARTICLE

Expert Systems WILEY

Effect of vocal tract dynamics on neural network-based speech recognition: A Bengali language-based study

Md Rakibul Hasan[1,2] | Md Mahbub Hasan[1] | Md Zakir Hossain[3,4]

[1]Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh

[2]Department of Electrical and Electronic Engineering, BRAC University, Dhaka, Bangladesh

[3]Agriculture and Food, Commonwealth Scientific and Industrial Research Organisation, Canberra, Australia

[4]Biological Data Science Institute, Australian National University, Canberra, Australia

Correspondence
Md Mahbub Hasan, Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology, Khulna 9203, Bangladesh.
Email: mahbub01@eee.kuet.ac.bd

Abstract

Although speech recognition has achieved significant success using integrated and efficient models, still some series of challenges remain as linguistic-acoustic patterns are perturbed by speakers' individual articulation gestures and environmental noises. Due to dynamic changes in the vocal tract cavity, word utterances yield temporal and perturbed linguistic-acoustic features, whereas vowel utterances yield less-perturbed quasi-stationary features. To recognize patterns as in vowels and words, the basic feedforward neural network (NN), among other methods, responds to these vocal tract-induced variabilities and has shown promising results because of its simple yet effective modelling of nonlinear data. We, therefore, present a comprehensive study on how these variabilities of acoustical features affect the speech token classification performances using NNs. We chose vocal tract resonance (formant frequency) as linguistic-acoustic feature. Our statistical evaluation of vocal tract-induced variabilities in seven Bengali vowels and words revealed that words have more variations than vowels. We used four-fold cross-validation in an NN with Adam optimizer to compute classification performances using five different metrics. Our experiments found that formant transitions and dispersions do not contribute to classification, and five-hidden-layered NN is optimum. In all different test cases, we justified our hypothesis—word classification falls behind vowel classification due to the variability induced by vocal tract dynamics. The optimum NN with 28,263 trainable parameters achieved the highest accuracy and AUC scores: 0.89 and 0.99 in vowels, and 0.64 and 0.91 in words.

KEYWORDS
Bengali speech classification, formant frequency, neural network, speech variability, vocal tract dynamics

1 | INTRODUCTION

Humans have God-gifted potential to recognize spontaneous or natural spoken languages with the highest accuracy, whereas implementing such capability in machines is not straightforward. In recent years, advanced models like neural networks (NNs) with necessary training data have made compelling achievements in less-natural read speech recognition technologies. However, similar accomplishments are not achieved in spontaneous speech recognition because non-linguistic spontaneous variables contaminate linguistic-acoustic variables, such as vocal tract resonance

Abbreviations: AUC-ROC, area under curve of receiver operating characteristic; CNN, convolutional neural network; COV, coefficient of variation; FN, false negative; FP, false positive; GMM, Gaussian mixture model; HMM, hidden Markov model; MFCC, Mel-frequency cepstral coefficient; NN, neural network; TN, true negative; TP, true positive.

frequency, fundamental frequency, and sound intensity (Deng & Ma, 2000; Deng et al., 2020). Non-linguistic variables originate from speakers' age, gender, emotion, region, culture, and vocal tract length (MacFarlane & Hay, 2015). Recognition of spontaneous, conversational speech requires consideration of internal structural information or generative mechanisms rather than only surface-level information. If only surface-level information is considered, theoretically infinite information is required to completely cover the overwhelming variability (Deng & Ma, 2000). Therefore, speech recognizer design must consider the variability induced by the vocal tract dynamics (Mitra et al., 2017).

Speech production overlaps comparatively stable vowels with dynamic consonantal articulation-related gestures of the vocal tract (Öhman, 1966, 1967). To infuse information in speech, the change of the vocal tract cavity with time is called vocal tract dynamics, which can be modelled by various time-varying filterings such as formant frequency, linear predictive coding, and Mel-frequency cepstral coefficient (MFCC). Formant frequencies refer to the resonant frequencies of the human vocal tract while producing speech. The formant trajectories of words are dynamic due to consonantal constrictions, whereas for vowels, these are quasi-stationary. These dynamic formant trajectories in words are more perturbed by the variations of speakers' vocal tract shapes than the quasi-stationary vowels. However, these dynamic articulatory activities are essential for conveying information to other humans and machines through speech tokens. In addition to the information transmission, these articulation dynamics can help describe neurological diseases such as Parkinson's disease (Gómez-Vilda et al., 2017) and Alzheimer's disease (Gosztolya et al., 2019). These diseases cause vocal tract variation that results in acoustical feature variation. Therefore, correct estimation of these vocal tract variations using classification performance might help diagnose these diseases.

In previous decades, most speech researchers employed statistical hidden Markov models (HMMs) to consider the temporal variability of speech and Gaussian mixture models (GMMs) to map the HMM states to the acoustic input features (Hinton et al., 2012; Jelinek, 1976; Young, 2008). Although GMM-HMM models have several advantages, NNs continuously replace their places by overcoming several significant shortcomings. In particular, HMM assumes the speech features as statistically independent, disregarding any possible correlation among individual features. Additionally, HMM strongly depends on the arbitrary assumption of probability density function associated with states (Trentin & Gori, 2003). GMM also has some disadvantages, including statistical inefficiency for nonlinear data space (Hinton et al., 2012). On the other hand, NN can successfully model nonlinear data (Hinton et al., 2012), and it is heavily used in pattern recognition tasks (Bishop, 2006; Looney, 1997). Several studies proved that NN with many hidden layers outperforms GMM-based models by a large margin in various speech recognition benchmarks (Hinton et al., 2012; Mohamed et al., 2012; Pan et al., 2012). Among several variants of NN architectures, feedforward NN is simple yet effective in pattern recognition tasks (Touvron et al., 2021). Moreover, other architectures—mostly derived from feedforward NN—might relegate vocal tract-induced variabilities, but the effect of these variabilities on speech token classification was the main focus of this article. We, therefore, selected feedforward NN to investigate the effects of vocal tract dynamics.

Through lifelong learning, humans adapt vocal tract-induced variabilities in recognizing speech. Based on the classification performance, *directions into velocities of articulators* (Guenther, 1994), *task dynamics* (Saltzman & Kelso, 1987), *state feedback control* (Houde & Nagarajan, 2011), and *feedback aware control of tasks in speech* models were formulated for computational speech-motor movement. These models are prominent in hearing impairment, stuttering, and phonatory learning. A well-trained NN as a speech token classifier considering perturbed formant can be integrated effectively into the acoustics to sound mapping tool in the auditory feedback module of these models.

According to Tripathi et al. (2020), vocal tract features achieve better speech classification than excitation source features. Although vocal tract dynamics play a significant role in speech classification, research reports, especially investigations on performance deviation due to vocal tract dynamics—either using GMM-HMM or NN classifier—are rarely found. Yang et al. (2020) and Sharmin et al. (2020) demonstrated speech token classification without explaining why such classification performances are achieved. Our previous work reported MFCC-based classification performance differences (Hasan & Hasan, 2020); this current article presents a comprehensive study on how vocal tract dynamics affect classification performance. We experimented with several variations in a feedforward NN model and input features (formant frequency). We, therefore, are also reporting the relative importance of formant frequency and its derived features (i.e., which features contribute more to speech recognition). This article discusses different NN configurations' relegation scenarios of vocal tract-induced acoustic perturbation. The main contribution of the articles is the numerical investigation of the effect of vocal tract dynamics on the NN-based speech classification performance. In this aspect, the present article shows how the classification performance varies with the number of NN layers and parameters. Secondly, considering Bengali pronunciation's regional diversity, we found the optimum NN structure for classifying speech tokens. Finally, we report that formant transitions and dispersions have no vital contribution to vowel and word classification, although canonical correlation-based studies reported their significance.

We chose Bengali because it is the sixth most spoken language globally, and it has 267.7 million total users worldwide, including 228.7 million native speakers (Eberhard et al., 2021). Popular speech-to-text recognizers (e.g., Google Assistant, Apple Siri, Microsoft Cortana, and Amazon Alexa) support Bengali either to some extent or not at all, which needs improvement for efficient real-time usage.

The paper is structured as follows. Related works are discussed in section 2. Starting from the data collection and ending with the NN model configuration, the whole methodology, including estimating dynamic acoustical variability, is illustrated in section 3. We then present the results of speech classification with particular emphasis on observed classification performance deviation in section 4. Lastly, we summarize the article with possible future works in section 5.

2 | LITERATURE REVIEW

The laryngeal energy triggers the resonance with a frequency below 5 kHz in most cases, so these first five formants are usually considered as speech features (Kent & Read, 2001). These hold crucial acoustic information, and accordingly, many researchers have been utilizing formant frequency to classify vowels and words. For example, Yan and Vaseghi (2003) used the first four formant frequencies to classify British, American, and Australian accents, reporting that formant frequency plays a significant role in classifying accents. Story and Bunton (2010) adopted the first three formant frequencies and their transitions in a study on articulation, revealing that these transitions contribute to the overall changes in the vocal tract shape at the time of speech production. They also provided direction that the time derivative of these transitions (basically the second derivative) would estimate the contributions of both vowels and consonants to the variation of vocal tract dynamics. Kent and Read (2001) emphasized formant transitions as an essential acoustic cue for speech perception, where second and third formant transitions are related to the place of production of a particular speech. Accordingly, exploring formant transitions (time derivative of formants) might be useful and, therefore, were utilized in our study.

Along with the five formants and their transitions, we used their dispersions (difference between formant pairs), which several studies also utilized as prominent features. Among these, López et al. (2013) used 12-dimensional feature vectors, and Hasan et al. (2015) used 9-dimensional feature vectors. These research groups employed the first five formants and four dispersions between two different formants pairs each time. Yusof and Yaacob (2008) used the first three formant frequencies and dispersions between them to classify Malaysian vowels, where they reported improvement for most of the vowels' classification when corresponding formant dispersions were incorporated. Some other studies also classified vowels by combining the three lowest formants and computed dispersions (Hillenbrand & Gayvert, 1993; Vuckovic & Stankovic, 2001).

Several notable works are available on speech classification; for example, Tripathi et al. (2020) classified speech mode (read and conversational) of four Indian languages—including Bengali—by employing vocal tract features in a single-layer feedforward NN model. Furthermore, they employed vocal tract features in multi-layer NN-based phone recognition. Yang et al. (2020) classified 10 English command words using three types of models, including feedforward NN and convolutional neural network (CNN). Similarly, Dawodi et al. (2020) classified 20 Dari speech tokens employing CNN. Saha et al. (2018) demonstrated improved speech classification employing vocal tract shape dynamics. There are several speech-token classification studies, particularly in the Bengali language. Sharmin et al. (2020) classified 10 Bengali spoken digits using CNN. Syfullah et al. (2018) classified 45 Bengali characters (vowels and consonants) using NN. Sumon et al. (2018) classified 10 isolated Bengali words using CNN. Mukherjee et al. (2018) classified seven Bengali vowels using a random forest algorithm. On top of these, Badhon et al. (2020) summarized several research reports on Bengali speech recognition.

Analysis of articulatory feature variations has several biomedical applications as it can help diagnose several diseases (Damper, 1982). Brabenec et al. (2017) utilized 14 combinations of speech tasks and acoustic features to diagnose Parkinson's disease. Hemmerling and Wojcik-Pedziwiatr (2020) predicted and estimated the same disease based on English vowel sounds. Therefore, proper analysis of articulatory feature variations is necessary for automatic speech recognition with biomedical applications. A unique application of NN-based consonant classification was demonstrated by Anjos et al. (2020), where a game was developed for children to practice pronouncing those difficult consonants to alleviate their pronunciation difficulties. Speech token recognition has several other applications: emotion recognition (Shahriar & Kim, 2019); voice assistance in smartphones, laptops, and vehicles; home automation (Yuksekkaya et al., 2006); and assisting physically-challenged and old-age people (Bolla et al., 2017; Qidwai & Shakir, 2012).

3 | METHODS

A sketch of the steps involved in analysing the effect of vocal tract dynamics and finding out the importance of different formant frequency features is depicted in Figure 1. The following subsections explain the steps in detail.

3.1 | Preparation of datasets

To evaluate the effect of vocal tract dynamic properties, we captured seven Bengali vowel sounds (/অ/[/ɔ/], /আ/[/a/], /ই/[/i/], /উ/[/u/], /ঋ/[/ri/], /এ/[/e/], and /ঐ/[/oi/]) and seven Bengali word sounds (/বোতল/[/bōtala/]{bottle}, /বন/[/bana/]{forest}, /কপি/[/kapi/]{copy}, /দোকান/[/dōkāna/]{shop}, /শেষ/[/śēṣa/]{end}, /সঠিক/[/saṭhika/]{correct}, and /উপরে/[/uparē/]{above}). The above texts within the square and the curly brackets indicate corresponding pronunciations and translations.

The Bengali language has 11 vowels. We selected seven of them, excluding two diphthongs and two longer versions of already selected two vowels, which have the same pronunciations in the present day speaking. We aimed to compare vowel and word classification performance, so we selected the same number of words. These seven specific words were selected because their pronunciation varies from speaker to speaker, and we were studying the effect of variations.

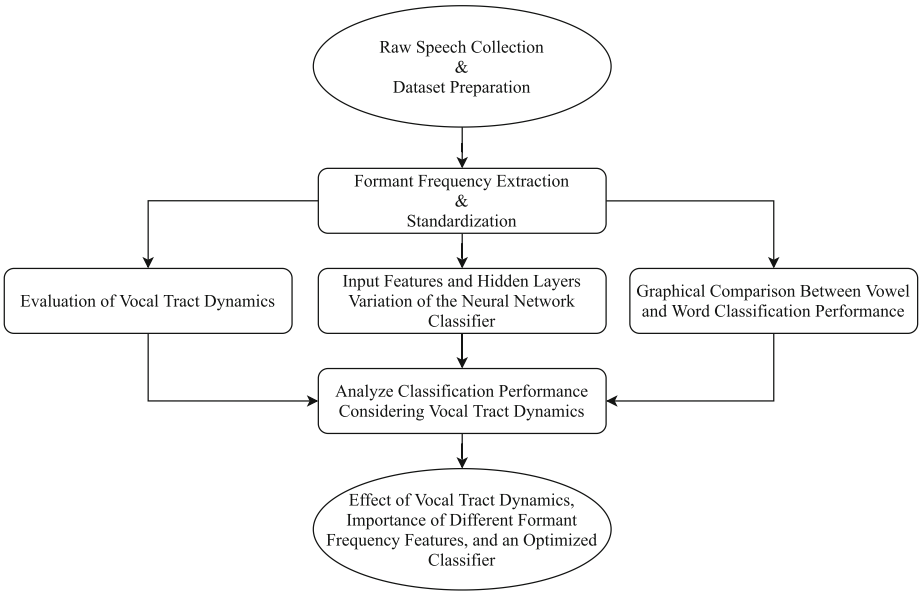


FIGURE 1 Steps involved in analysing the effect of vocal tract dynamics and finding out the importance of different formant frequency features

We selected 20 Bangladeshi speakers between the age range of 20–26 years whose native language is Bengali. The participants were university students who belong to different regions of Bangladesh. The Bengali language has several regional dialects that are more prominent than age group diversification. Thus, we diversified the dataset by including various regional volunteers. We guided the speakers to pronounce in two different accents—the normal one they use in their day-to-day life and any other regional Bengali accent they know about.

We recorded the sounds using the stereo-channel 'sound recorder' feature of a 'Xiaomi Redmi 3' smartphone at a sampling rate of 44,100 Hz in a normal environment (without noise suppressing environment). We then converted those stereo-channel sounds into mono-channel sounds using version 2.2.2 of the Audacity software (Audacity Team, 2018). While recording, we guided the speakers to speak all sounds in a continuous stream with a little pause between successive vowels/words so that we could easily separate those vowels/words later. We then clipped all classes of vowels and words using the same software by observing the waveform and hearing it simultaneously, and finally, we saved them as 32-bit float data-type in respective vowel/word classes. Eventually, we created both vowel and word datasets, with 40 utterances in each of the seven classes for both vowels and words. All data are publicly available for research purposes at Hasan and Hasan (2021).

3.2 | Feature extraction

We extracted five formant frequencies (F1–F5) for each of the seven vowels and words by sampling every 6 ms interval in a window length of 25 ms using a PRAAT script (Boersma & Weenink, 2001). These formant frequencies obtained at 6 ms intervals will be called formant "samples" throughout the manuscript. Additionally, four specific formant dispersions—the differences between a pair of formant frequencies—were calculated for all vowels and words according to Equation (1) to Equation (4).

$$F51 = F5 - F1 \tag{1}$$

$$F43 = F4 - F3 \tag{2}$$

$$F53 = F5 - F3 \tag{3}$$

$$F54 = F5 - F4 \tag{4}$$

Formant dispersion is how much a formant frequency deviates from another. For example, F51 is the positive difference between F5 and F1. The differences are always positive since the minuends are higher-order formant frequencies than the subtrahends. Subtraction was applied to each

formant sample of all vowels and words. If a specific vowel/word has M samples in each formant frequency (F1,F2,...F5), any formant dispersion of that vowel/word also has M values, as the above subtraction was applied in a sample by sample manner.

On top of these, we calculated five formant transitions across time by taking the second derivative with backward difference approximation given in Equation 5.

$$F_i'' = |\, F_i - 2F_{i-1} + F_{i-2} \,| \tag{5}$$

where i denotes a counter that goes from 1 to the total samples for each formant frequency. Applying the above equation to all five formants (F1 to F5), we extracted five formant transitions to be used as features.

3.3 | Statistical measures

To calculate the variation of a particular formant of a particular vowel or word, at first, we calculated mean (μ) according to Equation (6), standard deviation (σ) according to Equation (7), and coefficient of variation (COV) for a single sound file according to Equation (8).

$$\mu_j = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6}$$

$$\sigma_j = \frac{\sqrt{\sum_{i=1}^{n} (x_i - \mu_j)^2}}{n-1} \tag{7}$$

$$COV_j = \frac{\sigma_j}{\mu_j} \tag{8}$$

where x_i = formant samples, and n = total number of formant samples for the j^{th} sound file. Considering N number of sound files (N = 40) for that particular vowel or word, we calculated the overall COV of that particular formant frequency using Equation 9.

$$COV = \frac{\sum_{j=1}^{N} COV_j}{N} \tag{9}$$

3.4 | Neural network-based classification model

We utilized a feedforward NN architecture to classify quasi-stationary (vowels) and dynamic (words) speech tokens. Natural language processing tasks require recurrent neural networks (e.g., long short-term memory) or more advanced big transformer models (e.g., bidirectional encoder representations from transformers). It is worth mentioning that although we studied a natural language, our primary aim was to study the effect of vocal tract dynamics on speech token classification rather than conventional language processing tasks (e.g., next-word prediction or text summarization). Therefore, feedforward NN is more suitable than recurrent and transformer-based networks for our purpose.

We might have used CNN, but our primary aim might not be satisfied as CNN's convolution operation relegates vocal tract dynamics-induced variabilities (Hasan & Hasan, 2020). Moreover, feedforward NN is a part of CNN architecture: these are placed after the initial convolution layers for final prediction. Researchers are recently revisiting simpler feedforward NNs (also called multi-layer perceptrons) for classification tasks (Touvron et al., 2021). We, therefore, selected a basic feedforward NN model to study vocal tract dynamics. A similar study on speech classification (Tripathi et al., 2020) also used such a basic NN model. The following subsections explain the elements involved in the classification model.

3.4.1 | Feature standardization

Our extracted features had different ranges of values. Feature standardization or *Z*-score normalization transforms data to zero-mean and unit-variance, which helps NNs and other machine learning algorithms converge faster (Grus, 2019). We, therefore, standardized all acoustical features before feeding in their respective models according to Equation (10).

$$x'_c = \frac{x_c - \overline{x_c}}{\sigma_c} \tag{10}$$

Here, the subscript c indicates we performed the above operation on each feature, where the features are five formants, five dispersions, and four transitions. $\overline{x_c}$ and σ_c denote the mean and the standard deviation of that feature, and x'_c is the standardized features having a mean of zero and variance of one. We supplied these normalized features in the input layer of the feedforward NN-based classification model.

3.4.2 | Model configuration

The number of neurons in the input layer depends on the number of input features. The model's seven output neurons were utilized to represent seven speech classes: seven vowels in vowel classification and seven words in word classification. There could be one or multiple hidden layers between the input and output layers. The number of hidden layers and the number of neurons in these hidden layers cannot be defined by any hard and fast rule or formula. It varies based on the application area and can be best approximated by multiple experiments. We incremented the number of hidden neurons and layers from a smaller value and simultaneously tracked the model's performance. The performance plateaued after a specific number, which was optimum for our NN model.

We came up with five hidden layers as the optimum number in our experiments. Nevertheless, we are also reporting classifications using one, two, three, four, and six hidden layers to justify choosing five hidden layers. We also varied the number of input feature vectors in classification. We used four different sets of input features: only five input vectors (five formants), 10 input vectors (five formants and five dispersions), nine input vectors (five formants and four transitions), and 14 input vectors (all features—five formants, five dispersions, and four transitions). We, therefore, compared performances among combinations of five formant frequencies, five formant dispersions, and four formant transitions. The architecture of the feedforward NN model having five hidden layers and five formants in input, particularly for vowel classification, is shown in Figure 2.

Five neurons in the input layer were utilized to represent five formant frequencies (input feature vectors). When we employed all 14 feature vectors (five formants, five dispersions, and four transitions), there were 14 neurons in the input layer. Similarly, when it was a word classification model, the output classes were words rather than vowels shown in Figure 2.

The hyperparameters of the NN model, including the activation functions, optimizer, learning rate, loss function, batch size, and the number of epochs, were set through experiments. We tested with a specific setting and checked how it performed in terms of classification accuracy. Regarding the activation function, we came up with the *tanh* function for all hidden layers and the *Softmax* function for the output layer. After several experiments with different optimizers, including *Adadelta*, *SGD*, *RMSprop*, and *Adam*, we found *Adam* (Kingma & Ba, 2015) with a learning rate of 0.005 as the optimum optimizer. We used *categorical cross-entropy* as the loss function (Equation 11).

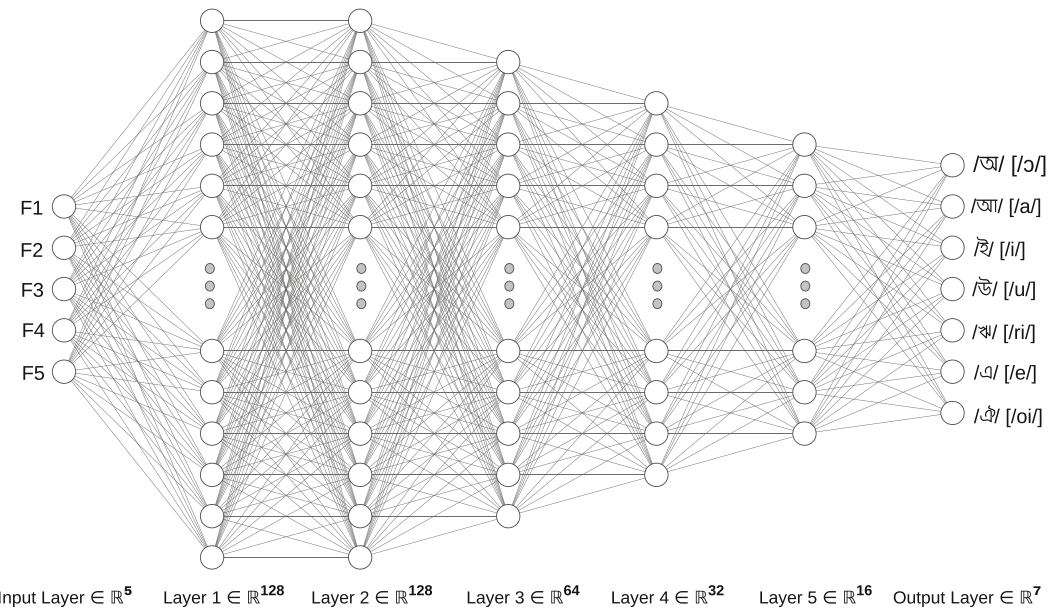


FIGURE 2 The architecture of the fully-connected feedforward neural network-based vowel classification model having five hidden layers. Input neurons represent the five formant frequency features, and output neurons represent the seven output vowel classes

$$\text{Loss} = -\sum_{j=1}^{\text{output size}} y_j \cdot \log \hat{y}_j \tag{11}$$

where y_j is the ground truth or target value, and $\hat{y}_j$ denotes the prediction made by the model for the j^{th} class.

During the training phase, example input–output patterns are served to the model so that it can tune its trainable parameters to predict the correct output class. Presentation of input features to the model can be either in sequential or batch mode. Sequential mode is also called stochastic mode, where individual training samples are passed from the input node to the output node one at a time, and then the trainable parameters are perturbed with respect to these individual sequences. It is optimal for larger datasets as it needs less computational power (Haykin, 2007). On the other hand, batch mode training involves all samples to be passed at once, and the parameters are accordingly perturbed based on all training samples as a whole. In our case, we chose batch mode training; the batch size was, therefore, the number of all training samples. After feature extraction, the vowel dataset had 14,161 formant samples, and the word dataset had 19,687 formant samples. As we used a four-fold cross-validation technique (discussed in section 3.4.3), the batch size was three-fourths of these total samples. One complete pass through the whole training dataset is called one epoch. We trained and validated our model for 300 epochs since the performance plateaued after this number of epochs.

Weights and biases are the two trainable parameters whose best values are searched throughout the NN training phase. The number of weights equals the number of connections between neurons of a layer and its preceding layer, whereas the number of bias parameters equals the number of neurons in that particular dense layer. Since we utilized a fully connected network, all input neurons were connected to all neurons of the first hidden layer, all neurons of the first hidden layer were connected to all neurons of the second hidden layer, and so on. The total weight parameter (say, W_l) in a dense layer (say, l) can be found by multiplying the number of neurons (say, n_l) by the number of input to that particular layer (i.e., the number of neurons in the preceding layer, n_{l-1}) shown in Equation (12). The total number of bias parameters (say, B_l) is equal to the number of neurons (n_l) of that particular layer l (Equation 13). Therefore, the total number of trainable parameters in a particular layer l can be derived according to Equation 14.

$$W_l = n_l \times n_{l-1} \tag{12}$$

$$B_l = n_l \tag{13}$$

$$\text{Thus, the total trainable parameters} = W_l + B_l \tag{14}$$

Particularly for the NN model shown in Figure 2, the numbers of neurons from the input layer up to the output layer were 5, 128, 128, 64, 32, 16, and 7, respectively. Thus, the total number of trainable parameters, according to Equation (14), was $(5 \times 128 + 128) + (128 \times 128 + 128) + (128 \times 64 + 64) + (64 \times 32 + 32) + (32 \times 16 + 16) + (16 \times 7 + 7) = 28{,}263$.

3.4.3 | Evaluation metrics

Some specific metrics are required to compare the performance between vowel and word classifications. We utilized five different metrics—confusion matrix, classification accuracy, area under the curve of the receiver operating characteristic (AUC-ROC), F_1 score, and Cohen's κ. These were guiding tools to attain optimum NN for classifying dynamic and quasi-stationary speech tokens (i.e., words and vowels).

A typical confusion matrix shown in Table 1 has four primary elements. True positive (TP) are actual positive examples predicted as positive, whereas false positive (FP) are actual negative examples predicted as positive. Similarly, true negative (TN) are actual negative predicted as negative, whereas false negative (FN) are actual positive cases predicted as negative by the classifier.

Several evaluation metrics can be defined from the confusion matrix (Davis & Goadrich, 2006), some of which are given in Equation (15) to Equation (18).

$$\text{Classification accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

TABLE 1 Elements of a confusion matrix

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

$$\text{True positive rate (Precision)} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{False positive rate} = \frac{FP}{FP + TN} \tag{17}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{18}$$

Classification accuracy measures the correctness of classification: it gives a ratio of correct predictions (both TP and TN) to all predictions the model made. Since the true label of the classes is known at both training and validation time, the algorithm compares the true label with the predicted label to provide the accuracy score. The accuracy score cannot explain the whole scenario when the number of samples varies among different classes. In that case, AUC-ROC measures the degree of separability among different classes. It is the area under the receiver operating characteristic curve—a plot of true-positive rate versus false-positive rate.

Precision or true-positive rate is the ratio of the number of correctly classified labels to all predictions the model picked as correct (including that are mistakenly predicted as correct); Recall or Sensitivity is the ratio of the number of correctly classified labels to all labels that should have been classified correctly (i.e., all ground truths). In most cases, an inverse proportional relationship exists between Precision and Recall. Therefore, a harmonic mean (Equation 19) of these two metrics—known as the F_1 score or F-measure—computes the model's performance better by giving a score proportional to the correct classification.

$$F_1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

Other than these derived metrics, the confusion matrix itself is used as another graphical metric that facilitates observing classification performance with respect to individual labels or classes.

Cohen's κ is a statistical metric that measures the inter-rater agreement, telling us how much the classifier performs other than a model that delivers just a random guess (Cohen, 1960). According to Landis and Koch (1977), a value of less than zero tells that the model gives a random guess; 0.81–1.00 implies an almost perfect model; 0.61–0.8 indicates a substantial-good model, and so on.

A higher value implies better classification for all evaluation metrics mentioned above. However, these metrics' computed results vary due to different reasons: random initialization of network parameters and data distribution between training and validation (i.e., what data belong to the training phase and what data belong to the validation phase). Both of these (parameter initializing and data splitting) were randomly decided in our experiments so that the classification performance could not be biased to initial parameters or input data distribution. The metrics' results, therefore, changed at different runs. To accommodate these changes and not be biased, we used four-fold cross-validation, a well-known strategy to combine performances of multiple runs in different settings (Refaeilzadeh et al., 2009).

4 | EXPERIMENTAL RESULTS AND DISCUSSION

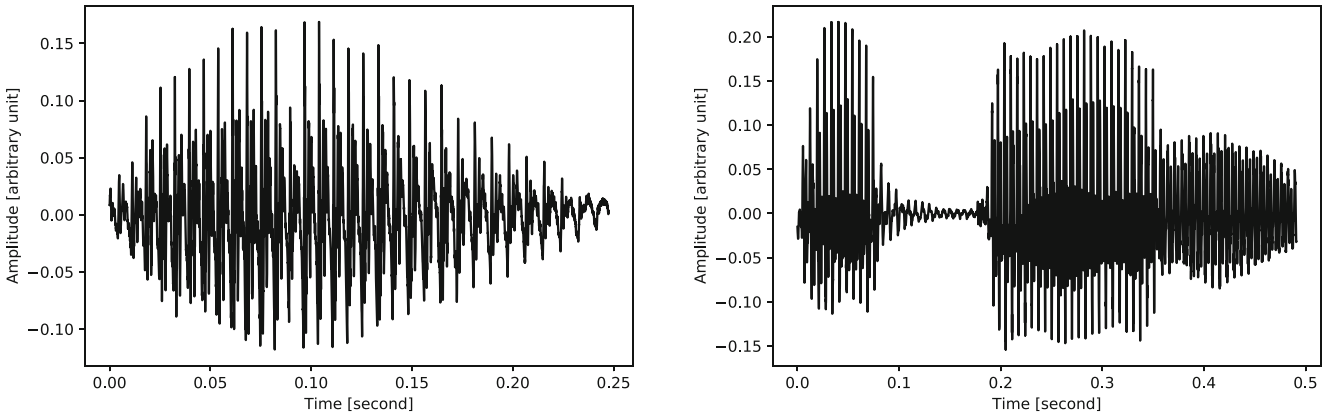
The waveforms of the / অ/[/ɔ/] vowel and the / বোতল/[/bōtala/] word from our captured data are shown in Figure 3a,b, respectively.

Figure 3 depicts that the vowel waveform was relatively steady, but the word waveform fluctuated as it (word) consists of vowels and consonants. The glottal pulse source energizes the quasi-stationary vocal tract while producing vowels, and so the vowel waveform becomes an approximately steady-state waveform. On the contrary, while producing words, consonantal constrictions induce the vocal tract's transitional nature, so the word waveform fluctuates (Figure 3b).

Formant trajectories usually illustrate the state of the transfer function of the vocal tract during speech production, and the dispersion of formant trajectories exhibits the numerical values of vocal tract dynamics (Mitra & Hasan, 2016). For visual comparison, the formant trajectories of the /অ/[/ɔ/] vowel and the /বোতল/[/bōtala/] word are shown in Figure 4a,b, respectively.

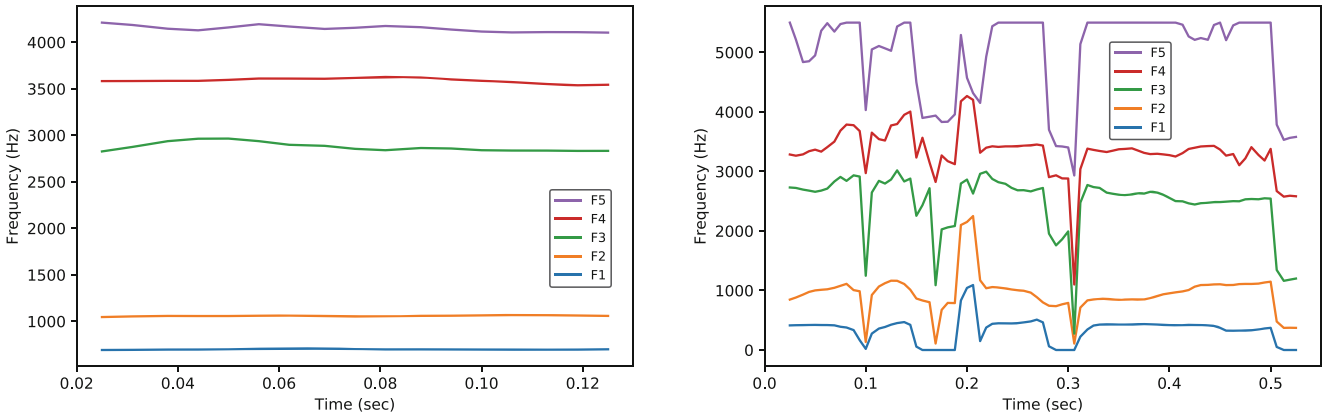
Between Figure 4a,b, a more dispersive nature was observed in words due to the presence of several consonantal constrictions during word production. These consonantal constrictions in the vocal tract induce dispersive behaviour in vocal tract resonance and filtering properties. To estimate the amount of vocal tract variation, we calculated the COV of all five formants of all vowels and words (Table 2).

The first formant had the highest dispersive nature, and the variation decreased for lower-order formants. Table 2 exhibits a sharp decrease in variation from F1 to F4 for both vowels and words. The variations between F4 and F5, especially for vowels, were less discernible. Words' formants had more variation than vowels, as proved by higher valued COVs.



(a) /অ/[/ɔ/] vowel (b) /বোতল/[/bōtala/] word

FIGURE 3 Waveshapes of a vowel and a word uttered by a particular speaker. The word waveshape shows more fluctuations than the vowel. Here, "arbitrary unit" refers to the normalized sound amplitude that represents the amount of air compression (above zero) or rarefaction (below zero) while producing the sound



(a) /অ/[/ɔ/] vowel (b) /বোতল/[/bōtala/] word

FIGURE 4 Comparison of formant trajectories between a vowel and a word. Word formants deviate more compared with vowel formants

4.1 | Input features and hidden layers variation

We classified both vowels and words by varying the number of input feature vectors and hidden layers. Table 3 presents four-fold cross-validated (average) scores with the standard deviations in parentheses.

Table 3 reports both vowel and word classification scores starting from a model of only one hidden layer. Having five formants in the input layer, the vowel classification model achieved an accuracy of 0.76 and an AUC-ROC score of 0.96. With one by one increase in the number of hidden layers, performance increased till the five-hidden-layered model. The low scores of models with less than five hidden layers indicate the network was under-fitted (since increasing the number of hidden layers to five increased the overall performance). Under-fitting refers that the number of parameters was insufficient to completely cover the acoustical variabilities. The five-hidden-layered model with five formant features in the input achieved the highest scores in the vowel classification. Moving to the six-hidden-layered model decreased vowel classification performance by 0.01 in all four metrics and increased word classification performance by 0.01 in two metrics (F_1 score and Cohen's κ). Therefore, considering a single optimum model for both vowel and word classifications, the five-hidden-layered model appeared optimum, achieving accuracy and AUC-ROC of 0.89 and 0.99 in vowel classification and 0.64 and 0.91 in word classification. Since the overall performance did not significantly increase in the six-hidden-layered model, the model parameters of the five-hidden-layered model were adequate to overcome the acoustical variabilities.

It is worthwhile to mention that increasing the number of layers also increases the total number of trainable parameters; additional parameters help the model understand input representations more thoroughly. Here is a caveat: more hidden layers increase the computational burden, a crucial limit in implementing speech recognition systems in low-resource edge devices.

TABLE 2 Coefficient of variation (COV) of the vowel and the word formant frequencies

	Coefficient of variation (COV)				
Speech	**F1**	**F2**	**F3**	**F4**	**F5**
/অ/[/ɔ/]	0.2959	0.1823	0.1064	0.0668	0.0603
/আ/[/a/]	0.3443	0.0896	0.0853	0.0602	0.0599
/ই/[/i/]	0.3288	0.2575	0.0670	0.0536	0.0669
/উ/[/u/]	0.4756	0.2167	0.1558	0.0672	0.0719
/ঋ/[/ri/]	0.4230	0.3123	0.1077	0.0829	0.0874
/এ/[/e/]	0.3340	0.2063	0.0659	0.0451	0.0647
/ঐ/[/oi/]	0.3915	0.4939	0.1312	0.0623	0.0701
/বোতল/[/bōtala/]	0.4623	0.3073	0.1633	0.1004	0.0869
/ বন/[/bana/]	0.6066	0.3503	0.2087	0.1226	0.1050
/কপি/[/kapi/]	0.7004	0.4955	0.1813	0.1027	0.0969
/দোকান/[/dōkāna/]	0.6829	0.3615	0.1936	0.1266	0.0926
/শেষ/[/śēṣa/]	0.9849	0.2661	0.1335	0.0847	0.0710
/সঠিক/[/saṭhika/]	0.9099	0.3926	0.1916	0.1133	0.0940
/উপরে/[/uparē/]	0.5334	0.4157	0.1753	0.1197	0.0932

Note: Higher values of COV denote higher variation in acoustical features (formant frequencies).

TABLE 3 Vowel and word classification performances in terms of cross-validated average scores (± standard deviation) at different numbers of feature vectors and hidden layers

Features	Layers	Parameters	Type	Accuracy	AUC-ROC	F_1 score	Cohen's κ
5F	HL_1	1671	Vowel	0.76 (± 0.00)	0.96 (± 0.00)	0.76 (± 0.00)	0.72 (± 0.00)
			Word	0.50 (± 0.01)	0.85 (± 0.00)	0.50 (± 0.01)	0.42 (± 0.01)
	HL_2	9479	Vowel	0.84 (± 0.00)	0.98 (± 0.00)	0.84 (± 0.00)	0.81 (± 0.00)
			Word	0.57 (± 0.00)	0.89 (± 0.00)	0.57 (± 0.00)	0.50 (± 0.00)
	HL_3	11,335	Vowel	0.86 (± 0.01)	0.98 (± 0.00)	0.86 (± 0.01)	0.84 (± 0.01)
			Word	0.61 (± 0.01)	0.90 (± 0.00)	0.60 (± 0.00)	0.54 (± 0.01)
	HL_4	11,751	Vowel	0.87 (± 0.01)	0.98 (± 0.00)	0.87 (± 0.01)	0.85 (± 0.01)
			Word	0.61 (± 0.01)	0.90 (± 0.00)	0.61 (± 0.01)	0.54 (± 0.01)
	HL_5	28,263	Vowel	**0.89** (± 0.00)	**0.99** (± 0.00)	**0.89** (± 0.00)	**0.87** (± 0.00)
			Word	**0.64** (± 0.01)	**0.91** (± 0.00)	0.63 (± 0.01)	0.57 (± 0.01)
	HL_6	32,423	Vowel	0.88 (± 0.00)	0.98 (± 0.00)	0.88 (± 0.00)	0.86 (± 0.00)
			Word	**0.64** (± 0.01)	0.91 (± 0.00)	**0.64** (± 0.01)	**0.58** (± 0.01)
5F, 5T	HL_5	28,903	Vowel	**0.89** (± 0.01)	0.98 (± 0.00)	**0.89** (± 0.01)	**0.87** (± 0.01)
			Word	0.63 (± 0.01)	0.90 (± 0.00)	0.63 (± 0.01)	0.57 (± 0.01)
5F, 4D	HL_5	28,775	Vowel	**0.89** (± 0.00)	0.98 (± 0.00)	**0.89** (± 0.00)	**0.87** (± 0.00)
			Word	**0.64** (± 0.01)	**0.91** (± 0.00)	**0.64** (± 0.01)	**0.58** (± 0.01)
5F, 5T, 4D	HL_5	29,415	Vowel	**0.89** (± 0.00)	0.98 (± 0.00)	**0.89** (± 0.00)	**0.87** (± 0.01)
			Word	**0.64** (± 0.00)	0.90 (± 0.00)	**0.64** (± 0.00)	**0.58** (± 0.00)

Note: Best scores for each classification appear in bold. 5F: Five formant frequencies; 5T: Five formant transitions; 4D: Four formant dispersions. HL_1: One hidden layer having 128 neurons. HL_2: Two hidden layers having 128 and 64 neurons, respectively. HL_3: Three hidden layers having 128, 64, and 32 neurons, respectively. HL_4: Four hidden layers having 128, 64, 32, and 16 neurons, respectively. HL_5: Five hidden layers having 128, 128, 64, 32, and 16 neurons, respectively. HL_6: Six hidden layers having 128, 128, 64, 64, 32, and 16 neurons, respectively. Parameters: Total number of trainable parameters in corresponding classification models.

Incorporating formant transitions and dispersions did not benefit classification: the performance scores were more or less equal in all four cases (all features, except dispersions, except transitions, and formants only) for the same five-hidden-layered configuration (Table 3). On average, classification performance was affected by the network's trainable parameters, not by the transitions and dispersions. However, canonical

correlation-based studies (Hasan et al., 2015; López et al., 2013) reported a significant influence of dispersions in classification. Our study proves no impact of transitions and dispersions on speech classification; the most plausible explanations are (1) the hidden layers and neurons extract the underlying dispersion and transitional relationship from the five formant features, and (2) the NN optimization algorithm further compensates the necessity of including the formant transitions and dispersions. This research, therefore, reveals that the number of hidden layers should be extended instead of incorporating formant transitions and dispersions.

4.2 | Graphical comparison between vowel and word classification

From our previous discussions, the optimum number of hidden layers appeared to be five, and formant dispersions and transitions appeared unnecessary. We, therefore, utilized that five-hidden-layered configuration with only formant frequencies to graphically illustrate training curves, validation curves, and performance comparisons between vowel and word classifications in terms of loss and accuracy (Figure 5).

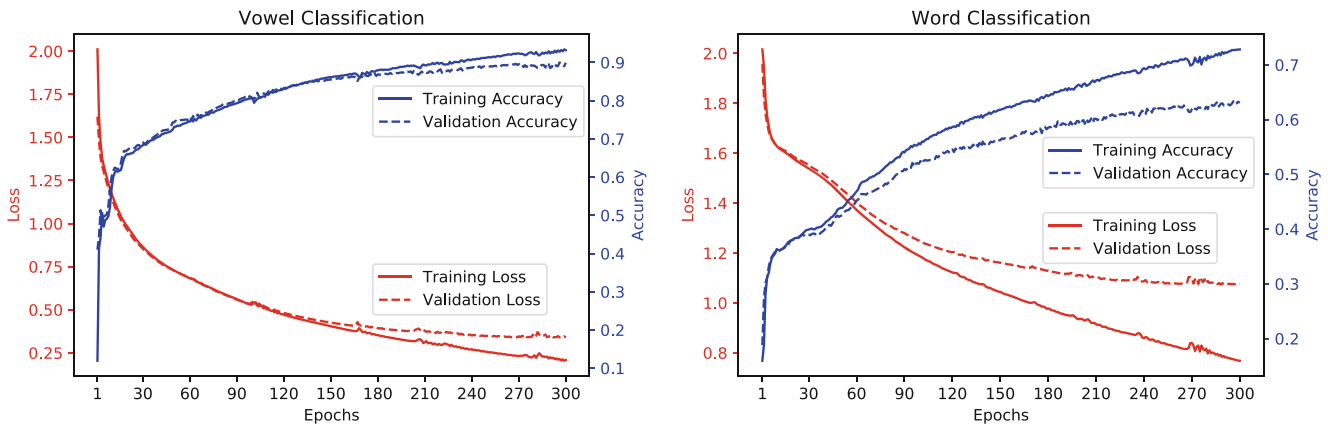
According to Stanford University's CS231n course materials,[1] a diffuse probability of $1/7$ is expected for each class since we had seven output classes. Thus, the initial loss is expected to be $-\ln(1/7) = 1.95$. Moreover, since the initial loss value came from NN parameters' random initialization, it closely varied around 1.95, as shown in Figure 5.

NNs tend to overfit due to higher variational training data (noise), which can be verified by observing training and validation curves: a considerable difference between training and validation scores implies overfitting, also known as high variance (Grus, 2019). As depicted in Figure 5a, training and validation curves had a good match in vowel classification, proving vowels have fewer higher variational acoustical features. On the contrary, a significant difference was observed between these curves in word classification (Figure 5b), proving words have higher variational acoustical features. Therefore, lesser overfitting in vowel classification indicates fewer random acoustical features, whereas higher overfitting in word classification indicates higher random acoustical features. The vocal tract produces these random acoustical features during word utterances, and thus, the information of dynamic acoustic features is not adequately modelled in word classification compared with vowel classification.

Figure 6 compares the validation phase's loss and accuracy curves between the vowel and the word classification. After 300 epochs, the validation loss went down to 0.3388 in vowel classification and 1.0757 in word classification. Similarly, the final validation accuracy was 0.8994 in vowel classification and only 0.6305 in word classification. A lower loss indicates better cost function minimization, revealing vowel classification is easier than word classification. The higher accuracy of vowel classification denotes better vowel classification performance than words.

Figure 7 depicts the confusion matrices, where top-left to bottom-right diagonal values represent the percentage of accurate classification for respective labels. It would have been the best model if these values were all one.

The confusion matrices exhibit that the model's correct prediction was above 0.90 except for /ঋ/[/ri/] and /ঐ/[/oi/]) vowels. We observed the best performance (0.97 accuracy) in vowel classification, particularly for the /আ/[/a/] vowel. The vowel classification model's worst prediction accuracy (0.79) was for the /ঐ/[/oi/]) vowel. On the contrary, the word classification model's best performance (0.82 accuracy) was for the /শষে/[/śēṣa/] word. In all other words, the model's performance was comparatively lower than vowel classification. The worst performance (only 0.45 accuracy) was found for the /উপরে/[/uparē/] word.



(a) Vowel classification

(b) Word classification

FIGURE 5 Loss minimization and accuracy score during training and validation. The differences between word's training and validation curves increase with increasing epochs, which denotes overfitting due to higher variational acoustic features

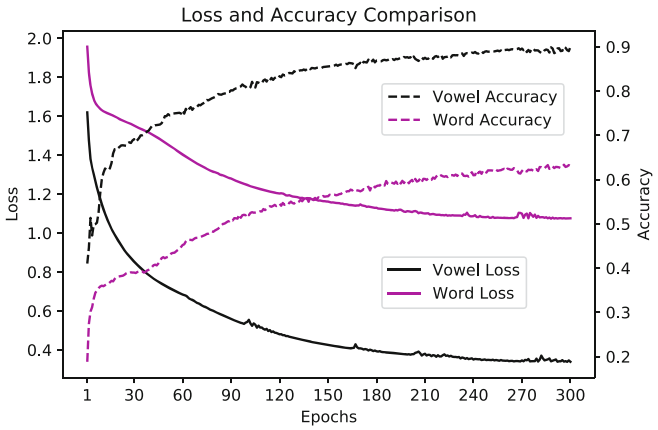
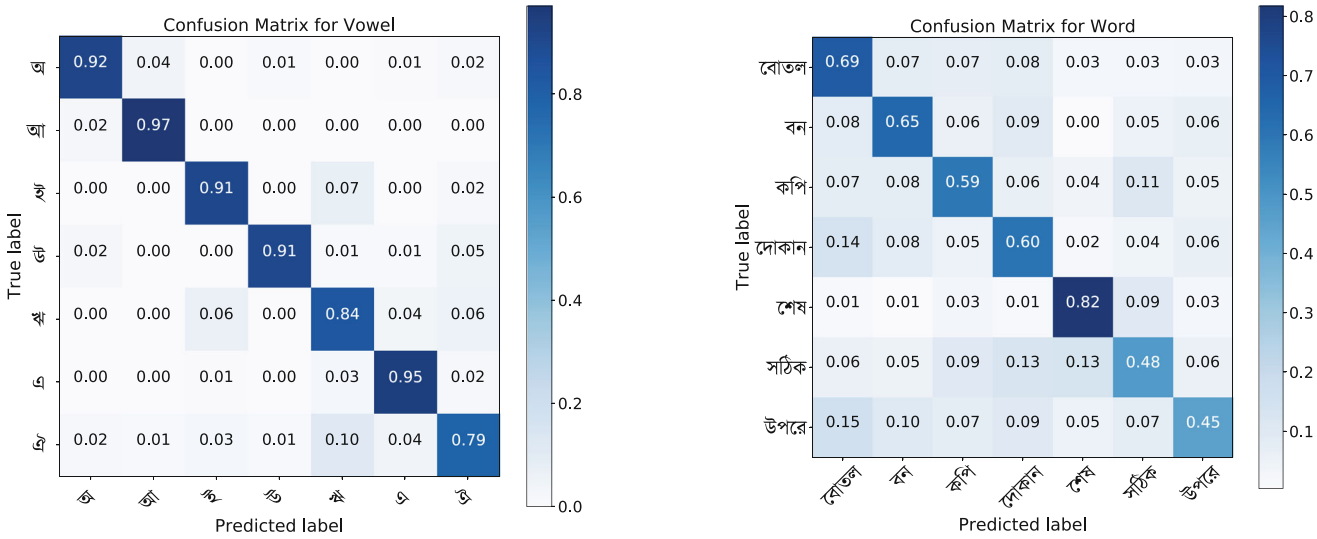


FIGURE 6 Validation loss and accuracy comparison between vowel and word classifications. Vowel loss is optimized to a smaller value than the word loss, and the vowel accuracy is higher than the word accuracy



a) Vowel classification

(b) Word classification

FIGURE 7 Confusion matrices for the vowel and the word classifications. The vowel classification model is quite suitable for correctly classifying vowels, but the word classification model is more confused to classify words correctly

4.3 | Summary and comparison

Word classification performance is significantly lower than vowel classification in all five performance metrics (Table 3). Additionally, word classification's deficiency is confirmed by comparing loss curves, accuracy curves (Figure 6), and confusion matrices (Figure 7). Such classification performance deviation lies in the unexplained variability of acoustical features. Vocal tract dynamics provide linguistic information related to dynamic acoustical features and random acoustic noises during word production. Additional parameterized NNs are required to accommodate these dynamic acoustic features in classifying and filtering individual articulation gestures of words and random acoustics (Table 3). Thus, lower parameterized NN poorly performs in word classification.

Tripathi et al. (2020) classified continuous speech into conversational and read modes and later developed a mode-specific phone recognition system. They reported 0.83 speech mode classification accuracy with vocal tract features. Yusof and Yaacob (2008) achieved 0.8704 accuracy in classifying five Malaysian vowels using NN with three raw formants and their dispersions. We should not strongly compare with these studies because the datasets and classification domains are different. However, this comparison proves the applicability of our NN-based classification model as we achieved as high as 0.89 classification accuracy for vowel classification. As Yusof and Yaacob (2008) showed reasonable accuracy with the first three formant frequencies and dispersions, more should be investigated on how vocal tract dynamics respond to such lower three formants and dispersions. Furthermore, we incorporated regional diversity in our speech data collection, lacking age-group diversity. A wider age range, especially data from younger and older people, would firmly support the finding.

5 | CONCLUSIONS AND FUTURE DIRECTION

The primary focus of this study was to investigate how NN-based speech recognition is affected by vocal tract dynamics, a crucial factor in spontaneous conversational speech recognition because of its variability. We separately classified seven Bengali vowels and seven Bengali words using two separate datasets. As acoustical features in the classifications, we chose formant frequency and its two derived feature sets: formant transitions and dispersions. A feedforward NN-based classification model was utilized with five well-known performance metrics—classification accuracy, AUC-ROC, F_1 score, Cohen's Kappa (κ), and confusion matrix—to evaluate the classification performance. Experimenting with different hidden layers and input features identified that five-hidden-layered NN is optimum, and the impacts of formant transitions and dispersions are insignificant. Between vowel and word classifications, word classification performance lagged by a large margin in all different tests. Vocal tract dynamics—determined from speech waveshapes, formant trajectories, and COVs—verified that words consist of more acoustic variability than vowels. Our vocal tract becomes relatively steady during vowel pronunciation, whereas it changes quite rapidly during word pronunciation due to coarticulation, which eventually induces acoustical feature variations. Therefore, the variation in words produced by vocal tract dynamics lowers the classification performance. Our NN-based speech classifier can be employed in computational speech motor movement models as an acoustic to sound mapping tool. Neurological disorders such as Parkinson's and Alzheimer's disease might be diagnosed by comparing patients' acoustic variabilities with the (regular Bengali speakers') variabilities presented in this article. Various speech dictation gadgets and services require proper detection of these isolated speech tokens, so service providers can serve Bengali language to their end-users with the aid of our presented Bengali speech token classification pipeline and the datasets. Formant-based features' relative importance revealed in this study will help select features in future research. Therefore, this study will help better design speech recognition-based devices and systems. However, this research did not quantify the amount of classification performance deviations. Future work might quantitatively relate vocal tract dynamics to classification performance deviations. Furthermore, since the data used in this study were collected from volunteers aged between 20 to 26, future work should include data from a wider age range to create more diversity.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ENDNOTE

[1] https://cs231n.github.io/neural-networks-3/#sanitycheck

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Mendeley Data at https://doi.org/10.17632/2h6975kdsx.1.

ORCID

Md Rakibul Hasan https://orcid.org/0000-0003-2565-5321
Md Mahbub Hasan https://orcid.org/0000-0003-2612-7248
Md Zakir Hossain https://orcid.org/0000-0003-1892-831X

REFERENCES

Anjos, I., Cavalheiro Marques, N., Grilo, M., Guimarães, I., Magalhães, J., & Cavaco, S. (2020). Sibilant consonants classification comparison with multi- and single-class neural networks. *Expert Systems*, *37*(6), e12620. https://doi.org/10.1111/exsy.12620

Audacity Team. (2018). *Audacity®: Free audio editor and recorder*. https://audacityteam.org/

Badhon, S. M. S. I., Rahaman, M. H., Rupon, F. R., & Abujar, S. (2020). State of art research in bengali speech recognition. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE. doi: https://doi.org/10.1109/ICCCNT49239.2020.9225650

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer-Verlag New York.

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9), 341–345.

Bolla, D. R., Shivashankar, Pavan, Ashwini, N. M., Kavya, V., & Mahesh, K. M. (2017). Voice enabled gadget assistance system for physically challenged and old age people. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)* (pp. 2081–2085). IEEE. doi: https://doi.org/10.1109/RTEICT.2017.8256966

Brabenec, L., Mekyska, J., Galaz, Z., & Rektorova, I. (2017). Speech disorders in parkinson's disease: Early diagnostics and effects of medication and brain stimulation. *Journal of Neural Transmission*, *124*(3), 303–334. https://doi.org/10.1007/s00702-017-1676-0

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Damper, R. I. (1982). Speech technology—Implications for biomedical engineering. *Journal of Medical Engineering & Technology*, *6*(4), 135–149. https://doi.org/10.3109/03091908209041006

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). Association for Computing Machinery. doi: https://doi.org/10.1145/1143844.1143874

Dawodi, M., Baktash, J. A., Wada, T., Alam, N., & Joya, M. Z. (2020). Dari speech classification using deep convolutional neural network. In *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1–4). IEEE. doi: https://doi.org/10.1109/IEMTRONICS51293.2020.9216370

Deng, L., & Ma, J. (2000). Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics. *The Journal of the Acoustical Society of America*, *108*(6), 3036–3048. https://doi.org/10.1121/1.1315288

Deng, Y.-C., Lin, C.-H., Liao, Y.-F., Wang, Y.-R., & Chen, S.-H. (2020). Prosodic information-assisted dnn-based mandarin spontaneous-speech recognition. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)* (pp. 134–138). IEEE. doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295010

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2021). *Ethnologue: Languages of the world* (24th ed.). SIL Internationalhttps://www.ethnologue.com

Gómez-Vilda, P., Mekyska, J., Ferrández, J. M., Palacios-Alonso, D., Gómez-Rodellar, A., & Rodellar-Biarge, V. (2017). Parkinson disease detection from speech articulation neuromechanics. *Frontiers in Neuroinformatics*, *11*(56), 1–17. https://doi.org/10.3389/fninf.2017.00056

Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & Hoffmann, I. (2019). Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features. *Computer Speech & Language*, *53*, 181–197. https://doi.org/10.1016/j.csl.2018.07.007

Grus, J. (2019). *Data science from scratch: First principles with python* (1st ed.). O'Reilly Media.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72*(1), 43–53. https://doi.org/10.1007/BF00206237

Hasan, M. M., Mitra, S. R., & Teramoto, K. (2015). Canonical correlation based impersonation quality determination algorithm for natural morphed speech. In *2015 IEEE International Conference on Telecommunications and Photonics (ICTP)* (pp. 1–4). IEEE. doi: https://doi.org/10.1109/ICTP.2015.7427943

Hasan, M. R., & Hasan, M. M. (2020). Investigation of the effect of mfcc variation on the convolutional neural network-based speech classification. In *2020 IEEE Region 10 Symposium (TENSYMP)* (pp. 1408–1411). IEEE. doi: https://doi.org/10.1109/TENSYMP50017.2020.9230697

Hasan, M. R., & Hasan, M. M. (2021). Isolated bengali vowel and word speech sounds. *Mendeley Data* doi: https://doi.org/10.17632/2h6975kdsx.1

Haykin, S. (2007). *Neural networks: A comprehensive foundation* (3rd ed.). Prentice-Hall.

Hemmerling, D., & Wojcik-Pedziwiatr, M. (2020). Prediction and estimation of parkinson's disease severity based on voice signal. *Journal of Voice*, *36*, 439. e9–439.e20. https://doi.org/10.1016/j.jvoice.2020.06.004

Hillenbrand, J., & Gayvert, R. T. (1993). Vowel classification based on fundamental frequency and formant frequencies. *Journal of Speech, Language, and Hearing Research*, *36*(4), 694–700. https://doi.org/10.1044/jshr.3604.694

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., … Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Houde, J., & Nagarajan, S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience*, *5*(82), 1–14. https://doi.org/10.3389/fnhum.2011.00082

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532–556. https://doi.org/10.1109/PROC.1976.10159

Kent, R. D., & Read, C. (2001). *The acoustic analysis of speech* (2nd ed.). Singular Publishing Group.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations*, San Diego. *arXiv* https://arxiv.org/abs/1412.6980

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174http://www.jstor.org/stable/2529310

Looney, C. G. (1997). *Pattern recognition using neural networks: Theory and algorithms for engineers and scientists*. Oxford University Press.

López, S., Riera, P., Assaneo, M. F., Eguía, M., Sigman, M., & Trevisan, M. A. (2013). Vocal caricatures reveal signatures of speaker identity. *Scientific Reports*, *3*(3407), 1–7. https://doi.org/10.1038/srep03407

MacFarlane, A. E., & Hay, J. (2015). Connecting linguistic variation and non-linguistic behaviour. *Linguistics Vanguard*, *1*(1), 259–270. https://doi.org/10.1515/lingvan-2015-1002

Mitra, S. R., & Hasan, M. M. (2016). Comparison of vocal-tract dynamics for bangla vowel and vowel-consonant-vowel sequence. In *International Conference on Advanced Information and Communication Technology* (pp. 1–7). https://www.ciu.edu.bd/icaict2016/publications/ICAICT-2016-Paper%20(6).pdf

Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C., Saltzman, E., & Tiede, M. (2017). Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. *Speech Communication*, *89*, 103–112. https://doi.org/10.1016/j.specom.2017.03.003

Mohamed, A.-R., Dahl, G. E., & Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 14–22. https://doi.org/10.1109/TASL.2011.2109382

Mukherjee, H., Phadikar, S., & Roy, K. (2018). An ensemble learning-based bangla phoneme recognition system using LPCC-2 features. In V. Bhateja, C. A. Coello Coello, S. C. Satapathy, & P. K. Pattnaik (Eds.), *Intelligent engineering informatics* (pp. 61–69). Springer Singapore. https://doi.org/10.1007/978-981-10-7566-7_7

Öhman, S. E. G. (1966). Coarticulation in vcv utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151–168. https://doi.org/10.1121/1.1909864

Öhman, S. E. G. (1967). Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, *41*(2), 310–320. https://doi.org/10.1121/1.1910340

Pan, J., Liu, C., Wang, Z., Hu, Y., & Jiang, H. (2012). Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why dnn surpasses gmms in acoustic modeling. In *2012 8th International Symposium on Chinese Spoken Language Processing* (pp. 301–305). doi: https://doi.org/10.1109/ISCSLP.2012.6423452

Qidwai, U., & Shakir, M. (2012). Ubiquitous arabic voice control device to assist people with disabilities. In *2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012)* (Vol. 1, pp. 333–338). IEEE. doi: https://doi.org/10.1109/ICIAS.2012.6306213

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of database systems* (pp. 532–538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Saha, P., Srungarapu, P., & Fels, S. (2018). Towards automatic speech identification from vocal tract shape dynamics in real-time mri. In *Proceedings of Interspeech 2018* (pp. 1249–1253). doi: https://doi.org/10.21437/Interspeech.2018-2537

Saltzman, E., & Kelso, J. (1987). Skilled actions: A task-dynamic approach. *Psychological Review*, *94*(1), 84–106. https://doi.org/10.1037/0033-295X.94.1.84

Shahriar, S., & Kim, Y. (2019). Audio-visual emotion forecasting: Characterizing and predicting future emotion using deep learning. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)* (pp. 1–7). IEEE. doi: https://doi.org/10.1109/FG.2019.8756599

Sharmin, R., Rahut, S. K., & Huq, M. R. (2020). Bengali spoken digit classification: A deep learning approach using convolutional neural network. *Procedia Computer Science*, *171*, 1381–1388. https://doi.org/10.1016/j.procs.2020.04.148

Story, B. H., & Bunton, K. (2010). Relation of vocal tract shape, formant transitions, and stop consonant identification. *Journal of Speech, Language, and Hearing Research*, *53*(6), 1514–1528. https://doi.org/10.1044/1092-4388(2010/09-0127)

Sumon, S. A., Chowdhury, J., Debnath, S., Mohammed, N., & Momen, S. (2018). Bangla short speech commands recognition using convolutional neural networks. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)* (pp. 1–6). IEEE. doi: https://doi.org/10.1109/ICBSLP.2018.8554395

Syfullah, S. M., Zakaria, Z. B., Uddin, M. P., Rabbi, M. F., Afjal, M. I., & Nitu, A. M. (2018). Efficient vector code-book generation using k-means and linde-buzo-gray (LBG) algorithm for bengali voice recognition. In *2018 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)* (pp. 1–4). IEEE. doi: https://doi.org/10.1109/ICAEEE.2018.8642994

Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., … Jégou, H. (2021). Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv*https://arxiv.org/abs/2105.03404

Trentin, E., & Gori, M. (2003). Robust combination of neural networks and hidden markov models for speech recognition. *IEEE Transactions on Neural Networks*, *14*(6), 1519–1531. https://doi.org/10.1109/TNN.2003.820838

Tripathi, K., Reddy, M. K., & Rao, K. S. (2020). Multilingual and multimode phone recognition system for indian languages. *Speech Communication*, *119*, 12–23. https://doi.org/10.1016/j.specom.2020.02.006

Vuckovic, V., & Stankovic, M. (2001). Formant analysis and vowel classification methods. In *5th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service. TELSIKS 2001 (cat. No. 01ex517)* (Vol. 1, pp. 21–24). IEEE. doi: https://doi.org/10.1109/TELSKS.2001.954841

Yan, Q., & Vaseghi, S. (2003). Analysis, modelling and synthesis of formants of british, american and australian accents. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP '03)* (Vol. 1, pp. 712–715). IEEE. doi: https://doi.org/10.1109/ICASSP.2003.1198880

Yang, X., Yu, H., & Jia, L. (2020). Speech recognition of command words based on convolutional neural network. In *2020 International Conference on Computer Information and Big Data Applications (CIBDA)* (pp. 465–469). IEEE. doi: https://doi.org/10.1109/CIBDA50819.2020.00110

Young, S. (2008). Hmms and related speech recognition technologies. In J. Benesty, M. M. Sondhi, & Y. A. Huang (Eds.), *Springer handbook of speech processing* (pp. 539–558). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-49127-9_27

Yuksekkaya, B., Kayalar, A. A., Tosun, M. B., Ozcan, M. K., & Alkar, A. Z. (2006). A gsm, internet and speech controlled wireless interactive home automation system. *IEEE Transactions on Consumer Electronics*, *52*(3), 837–843. https://doi.org/10.1109/TCE.2006.1706478

Yusof, S. A. M. M. P., & Yaacob, S. (2008). Classification of malaysian vowels using formant based features. *Journal of Information and Communication Technology*, *7*, 27–40.

AUTHOR BIOGRAPHIES

Md Rakibul Hasan currently serves as a faculty member in the Electrical and Electronic Engineering Department at BRAC University, Dhaka, Bangladesh. He received his BSc (2019) and MSc (2021) degrees in Electrical and Electronic Engineering from Khulna University of Engineering & Technology, Khulna, Bangladesh. Along with teaching, he strongly focuses on active research with global collaboration. In particular, his research interests include applied machine learning, signal processing, IoT, and embedded systems. He worked on deep learning applications in Bengali speech recognition in his Bachelor's and Master's thesis. As a young academician, he has several peer-reviewed publications related to applied deep learning and machine learning.

Md Mahbub Hasan was born in Jashore, Bangladesh, in 1982. He received the BSc (2006) and MSc (2010) from Khulna University of Engineering & Technology, Khulna, Bangladesh, and PhD (2013) from Saga University, Japan, all in Electrical and Electronic Engineering. Currently, he is serving as a faculty member in the Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh. His research interests involve multidimensional signal processing and mathematical modelling. He published articles in the field of speech processing and optical system modelling. He is a member of IEEE and the Institute of Engineers Bangladesh.

Md Zakir Hossain completed the BSc (2011) and MSc (2014) from Khulna University of Engineering & Technology (KUET), Khulna, Bangladesh in Electrical and Electronic Engineering, and PhD (2019) from Australian National University (ANU), Canberra, Australia in Computer Science. He has working experience with several universities and organizations including KUET, ANU, University of Canberra, Macquarie University, and CSIRO. His research direction leads to develop advance technologies for diagnosing and managing diseases by analysing sound signals, physiological signals, and multi-omics data including speech and cough sounds. He has published a number of articles in the field of affective computing and computer vision.

How to cite this article: Hasan, M. R., Hasan, M. M., & Hossain, M. Z. (2022). Effect of vocal tract dynamics on neural network-based speech recognition: A Bengali language-based study. *Expert Systems*, *39*(9), e13045. https://doi.org/10.1111/exsy.13045