T.R.

GEBZE TECHNICAL UNIVERSITY

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

NASAL VOICE DETECTOR WITH SPEECH THERAPISTS

RAHMET ALI ÖLMEZ

SUPERVISOR
PROF. DR. YUSUF SINAN AKGÜL

GEBZE
2022

T.R.
GEBZE TECHNICAL UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

NASAL VOICE DETECTOR WITH SPEECH THERAPISTS

RAHMET ALI ÖLMEZ

SUPERVISOR
PROF. DR. YUSUF SINAN AKGÜL

2022
GEBZE

GEBZE TECHNICAL UNIVERSITY	GRADUATION PROJECT JURY APPROVAL FORM

This study has been accepted as an Undergraduate Graduation Project in the Department of Computer Engineering on 31/08/2021 by the following jury.

JURY

Member
(Supervisor) : Prof. Dr. Yusuf Sinan Akgül

Member : Dr. Yakup Genç

ABSTRACT

Hypernasality is a resonance disorder seen in children and adults. A hypernasal voice is produced when too much air escapes through the nose while speaking. Some of the cases of hypernasality can only be cured with surgery. However there are also many cases that are curable with speech therapy. This disorder can be diagnosed either by a special device that called the nasality microphone where the resonance of the nasal and oral area are compared, or by a professional that listens to the speech of the patient. The aim of this study is to do the analysis of detecting nasal voices automatically with the help of deep learning techniques and help children with hypernasality to improve their normal speaking voice, with the help of a game.

Keywords: nasality, speech therapy.

ÖZET

Hipernazalite çocuklarda ve yetişkinlerde görülen bir rezonans bozukluğudur. Hipernazal sesler, konuşurken burundan gereğinden fazla hava çıkmasıyla oluşur. Hipernazalite vakalarından bazıları sadece ameliyat ile iyileştirilebilmektedir. Ancak, konuşma terapisiyle düzelebilen vakalar da vardır. Bu bozuklukluğun teşhisi nazalite mikrofonu denilen ve burun ve ağız boşluklarındaki rezonansın karşılaştırıldığı özel bir alet ile ya da hastanın sesini dinleyen bir uzman tarafından yapılabilmektedir. Bu çalışmanın amacı, nazal ses analizini derin öğrenme tekniklerini kullanarak otomatik yapmak ve hipernazaliteye sahip çocukların normal konuşma seslerini geliştirmeye bir oyun yoluyla yardımcı olmaktır.

Anahtar Kelimeler: nasality, speech therapy.

ACKNOWLEDGEMENT

I would like to thank my supervisor Prof. Dr. Yusuf Sinan Akgül for guiding me throughout the project. I would also like to thank Mr. Ilhan Aytutuldu for his invaluable help and every person that contributed to this project by sharing their vocal data.

Rahmet Ali Ölmez

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol or **Abbreviation**	:	**Explanation**
MFCC	:	Mel Frequency Cepstral Coefficients
CNN	:	Convolutional Neural Network
LSTM	:	Long Short Term Memory

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Hypernasality is a resonance disorder seen in children and adults. A hypernasal voice is produced when too much air escapes through the nose while speaking. Some of the cases of hypernasality can only be cured with surgery. However there are also many cases that are curable with speech therapy. This disorder can be diagnosed either by a special device that called the nasality microphone where the resonance of the nasal and oral area are compared, or by a professional that listens to the speech of the patient. In this project, our objective is to do the analysis of detecting nasal voices automatically with the help of deep learning techniques including CNN and LSTM networks and provide a game for children with hypernasality to help them easily improve their normal speaking voice.
In this report we will present our project architecture, experiment results and implementation details.

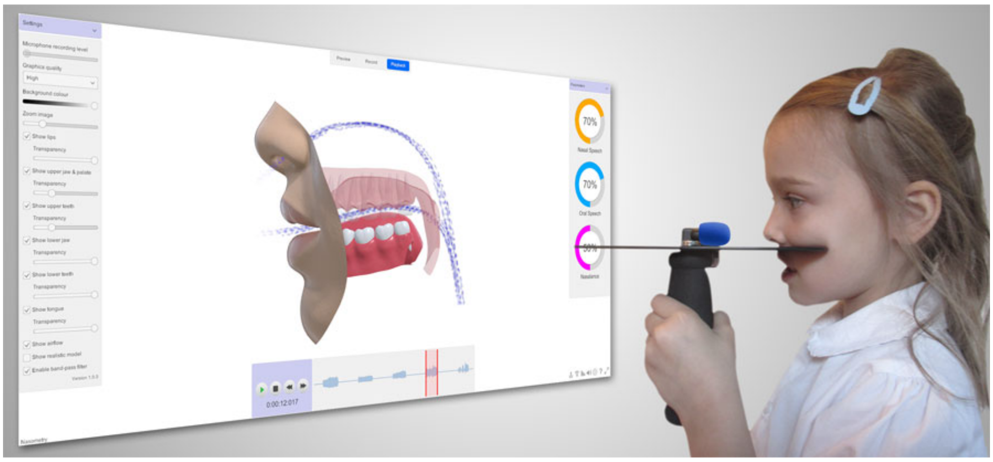


Figure 1.1: The nasality microphone in use. [16]

2. LITERATURE REVIEW

2.1. Nasal Speech Detection

Many studies have been made in order to be able to distinguish nasal speech from normal speech. One of the studies uses a support vector machine method using features of vowel space area and Mel-frequency cepstral coefficients. In this study, the hypernasality detection scores an accuracy of 86.89% for sustianed vowels and up to 91.70% for vowels in contexts of high pressure consonants. [17]

In another research, hypernasality was estimated by calculating a quantity from comparing the distance between the sequences of cespstrum coefficients extracted from Autoregressive model and Autoregressive Moving Average model. [18]

A study done in 2019 uses LSTM-DRNN (Long Short-Term Memory based Deep Recurrent Neural Network) to detect and diagnose hypernasal speech seen in children with cleft palate disease. The trained model reaches an accuracy of 93.35%, which is a higher detection accuracy than shallow classifiers. This study concludes that there is a potential of deep learning on pathologist speech detection. [19]

Convolutional Nerual Networks (CNN) have also been trained previous studies. One of them describes a feature-independant ent-to-end algorithm that uses a CNN to detect hypernasality. The input of the model used in this study is a speech spectrogram. This detection algorithm scores an F1-score of 0.9485 using a dataset consisting of children's speech, and 0.9746 using a dataset that is created from audults' speech. [20]

2.2. Speech Therapy Games

Many speech therapy games have been studied and developed that aim to help children improve their speech in an entertaining manner. Some examples include: Apraxiaville [21], Articulation Station [22], ArtikPix [23], Tabby Talks [24], Tiga Talk [25], Pocket SLP [26]. Most of these games do not provide feedback on the users' speech. Tabby Talks [24] is a mobile game that does provide feedback, by using an automatic speech recognition engine that runs on a remote server. This engine scores the speech productions the users make.

In a research made in 2017 [27], a focus group with cleft lip specialists, researchers and game developers was held. In the focus group, researchers and game developers agreed that maximizing immersion through emotionally motivating elements would best hide the repetitive speech exercises. In their game design, they validate the following aspects of their game are are emotionally motivating:

- Characters that are relatable and have the capacity to create empathy
- An overarching plot for our characters that is interesting and defines the player's goals
- Narrators that engage and encourage the player to continue helping the game characters
- A Procedural Content Background Art Generator that creates rich environments to experience
- Mechanics that seamlessly incorporate conversational speech recognition into completing game objectives

3. PROJECT ARCHITECTURE

This project has two main modules:

- Detection Module: This module takes the users' audio data and outputs either nasal or non-nasal.
- Game Module: The game module takes the outputs of the detection module and uses them for the movements and logic of the game.

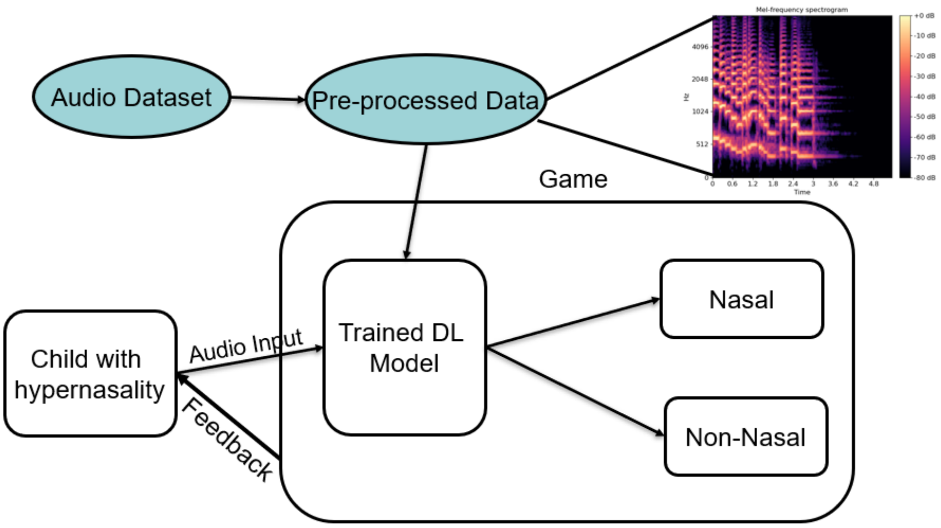


Figure 3.1: Diagram of the project [28]

As seen in Figure 3.1, the detection module, which consists of a trained deep learning model, takes the user's speech as input. In the module, the audio data is pre-processed to match the model's parameters. Then the module outputs either nasal or non-nasal. The game module uses the information to provide feedback to the user.

4. DATASET

4.1. Data Collection Method

The data we needed for this project where true nasal voices were recorded was unfortunately not easy to obtain. Since we needed hours of data, we collected the vocal data by ourselves by attempting to approximate our voices to nasal sounds by closing the nostrils of our nose with our hands. We read texts that were provided by speech therapists.

4.2. Audio Format

The sampling rates of the recordings were 44100 Hz with a quantization of 16 bits. The audio files were processed to be single channel, making the sampling rates 22050 Hz. The files were recorded in .WAV format.

4.3. Quantity of the Data

At first, two hours of audio was recorded (by one person) and another one hour was added to the data, recorded by five different people.

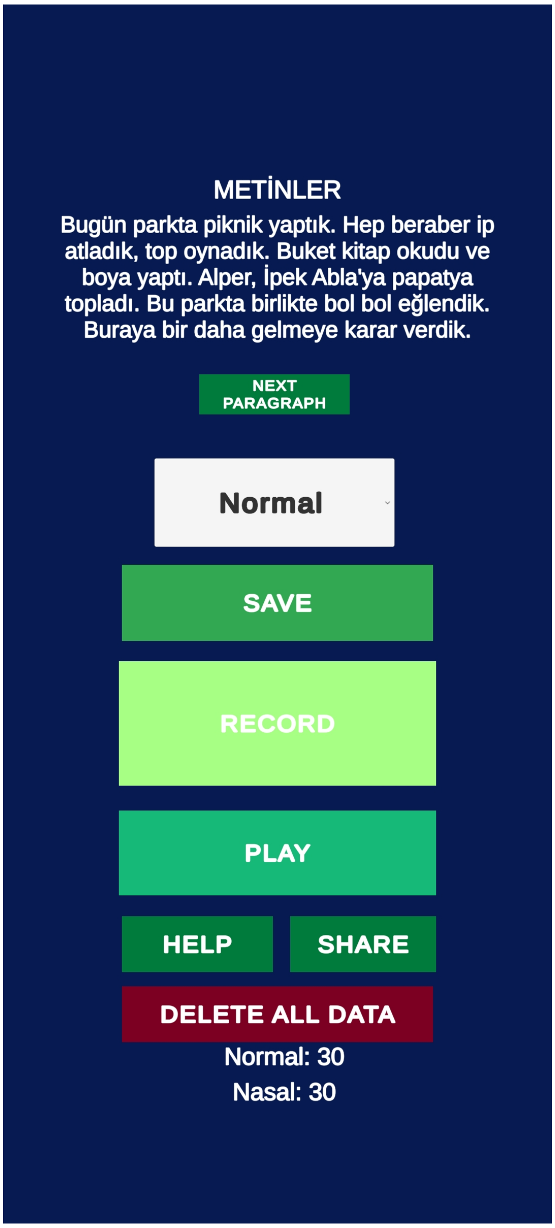
4.4. Easing Data Collection

Recording our voice is not very hard using a regular smartphone for the first few times, but the process gets tedious quite quickly; especially when naming the audio files that were recorded, which also takes additional time.

To try to solve this problem, we developed a mobile application that does the naming, compressing the audio files and sharing automatically. The application also includes the texts to be read while recording data. See figure 4.1 for the screenshot of the application.

4.5. Pre-Processing Data

The raw data had to be processed to match the parameters of the deep learning model. The models we trained take MFCC's as input, which are extracted features of spectrograms. Spectrograms basically are image representations of audio data. We are specifically using mel-spectrograms which are good at capturing the characteristics of phonetic speech. [29]



METİNLER

Bugün parkta piknik yaptık. Hep beraber ip atladık, top oynadık. Buket kitap okudu ve boya yaptı. Alper, İpek Abla'ya papatya topladı. Bu parkta birlikte bol bol eğlendik. Buraya bir daha gelmeye karar verdik.

Normal: 30

Nasal: 30

Figure 4.1: A screenshot of the application developed for easier data collection.

5. METHOD

5.1. Training Deep Learning Models

In this study we have trained two types of deep learning models and tried to increase their accuracy. The first one is a CNN (convolutional neural network) and the other is an LSTM (long short term memory) network.

5.1.1. CNN

The CNN model consists of three convolutional layers, each followed by a max pool layer. The convolutional layers use *rectified linear unit* as their activation functions and the output layer uses *softmax*. Figure 5.1 shows the details of the architecture.

5.1.2. LSTM

The LSTM model has two LSTM layers, the first being a *sequence to sequence layer*, and a *dense layer*. This model also uses *softmax* in the output layer. The details of the architecture can be seen in figure 5.2.

5.1.3. Test Results

After collecting two hours of data which as stated earlier, was the data of a single person. The accuracies were 0.99, where the test data also belonged to the same person. After testing the model with different people, the accuracy significantly dropped, which means that the model was overfitted. To solve this problem, an extra hour of audio data was added to the dataset. The test results for the two deep learning models with the final dataset are shown in table 5.1

Table 5.1: Comparison of test results.

Model	Accuracy	F-Measure
CNN	0.976	0.978
LSTM	0.977	0.977

As seen in the table, the test results are still very high despite the increased number of people that contributed to the dataset.

We have also evaluated the models with speech data of people they have never been

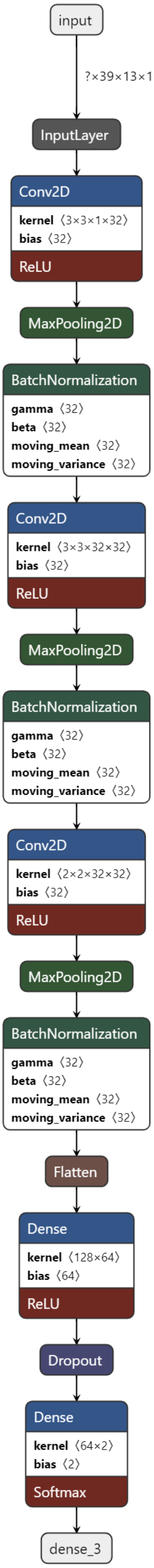


Figure 5.1: Architecture of CNN model.

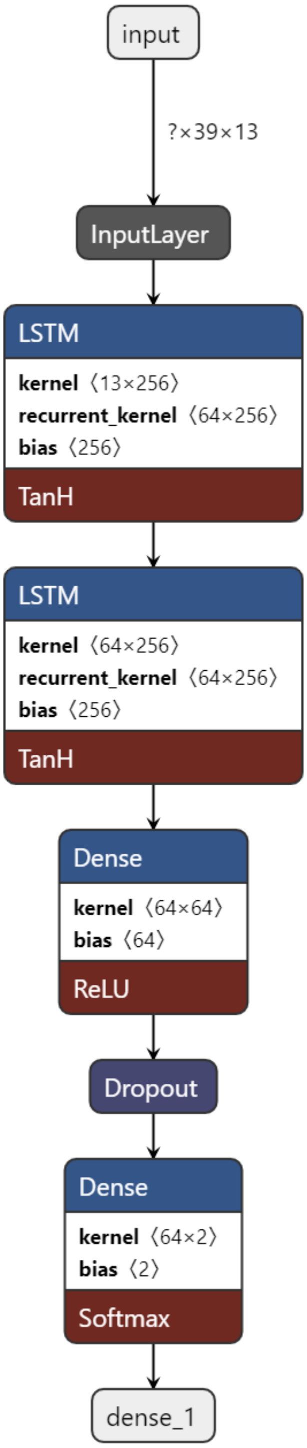


Figure 5.2: Architecture of LSTM model.

trained with. It is worth to note that this dataset had a size of about six minutes. The results are shown in table 5.2

Table 5.2: Accuracy scores of models for data of people they have not been trained with. (LSTM was only trained with 3 hours of data)

Model	2 Hour Data	3 Hour Data
CNN	0.581	0.410
LSTM	-	0.521

5.1.4. Training Environment

To be able to train the models faster, we have used Google Colab as our development platform.

5.2. Speech Therapy Games

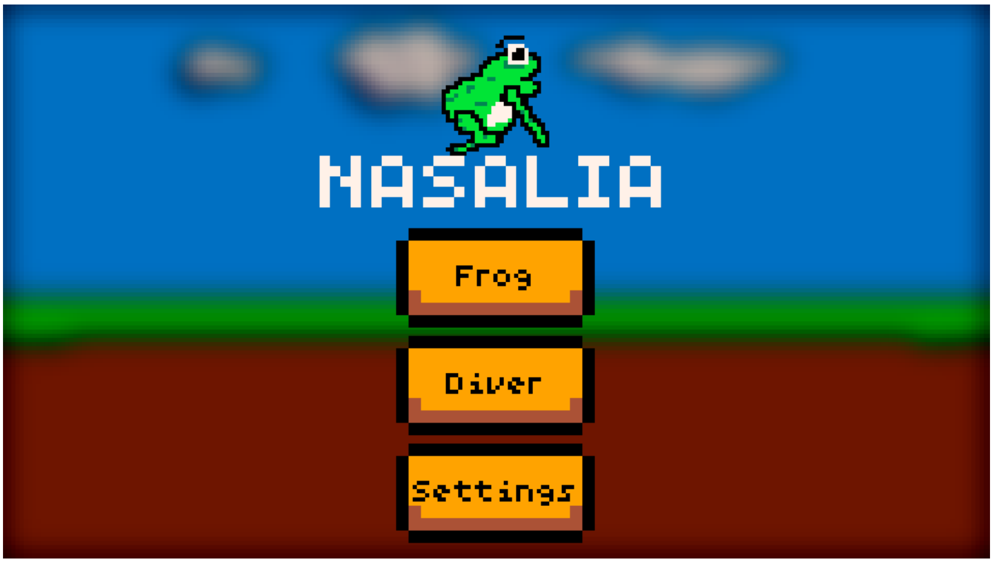


Figure 5.3: Main menu of the game.

The aim for the speech therapy games were to help children with the disorder to get better in normal speaking by helping them clearly see when they are speaking in a nasal voice and when in a normal one. To achieve this we made the games quite easy to understand and with minimal game mechanics.

5.2.1. Game Themes

The game theme ideas are taken from Rose Medical - Speech Therapy Software and Instrumentation. [30]

The two mini-games are as follows:

- Frog: The goal is to catch the fly using the frog's controllable tongue.
- Diver: The goal is to help the diver and make him reach the treasure chest.

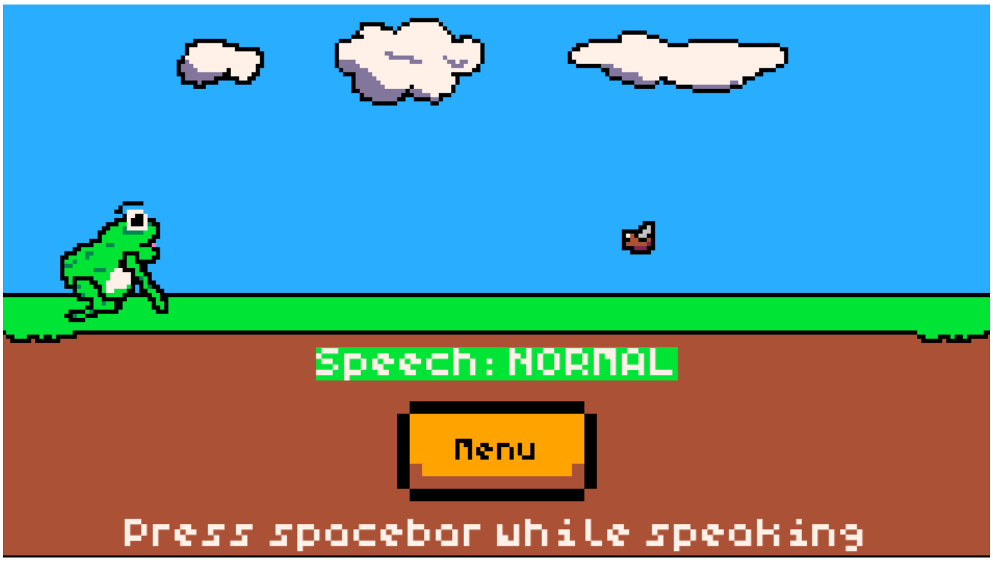


Figure 5.4: The frog game.

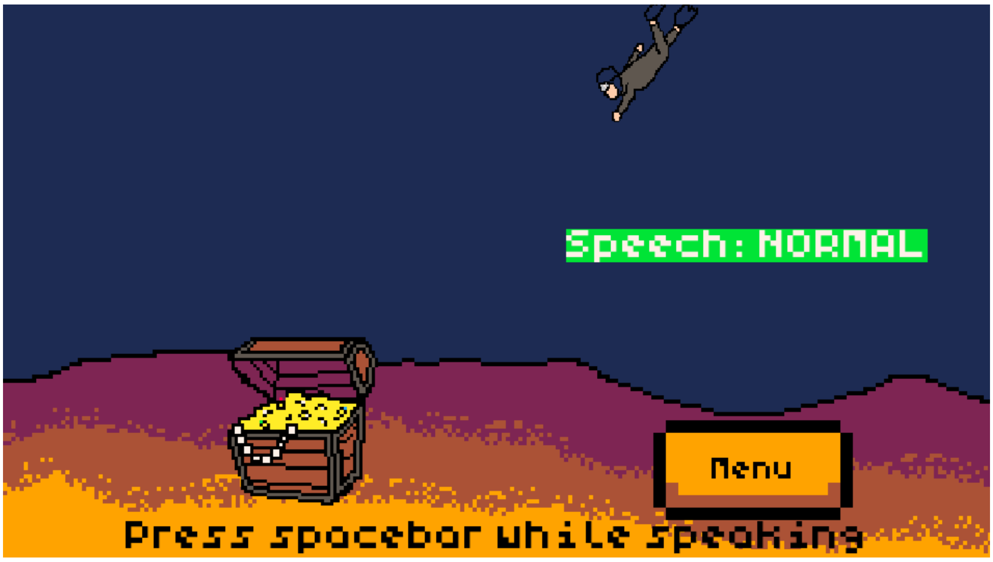


Figure 5.5: The diver game.

5.2.2. Game Logic

As we have previously stated, the game logic is simple. The voice of the users only control one object in the game. That object makes positive progress if the user talks in a normal voice, and negative progress for nasal voice. This object is the tongue of the frog in the frog game, and the diver in the diver game.

5.2.3. Difficulty Level

The frog game's difficulty level is designed to be adjustable. This is accomplished by making the fly to be caught a controllable object. As the fly is moved away from the frog, the user has to speak in a normal voice, more than before. Figures 5.6 and 5.7 show how difficulty levels can be adjusted.

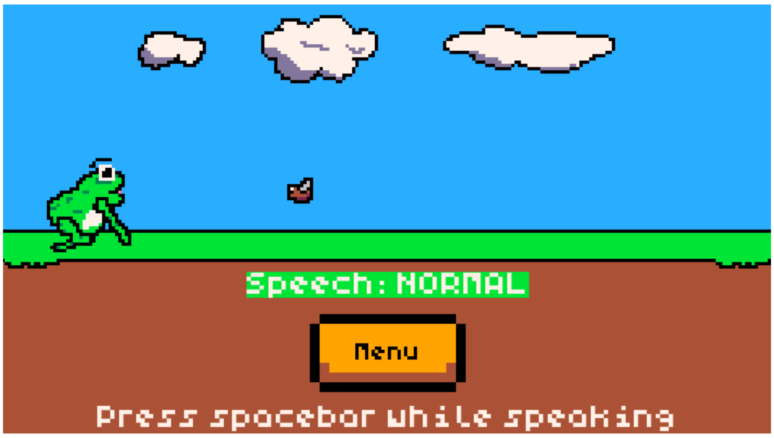


Figure 5.6: Decreased difficulty, achieved by moving the fly towards the frog.

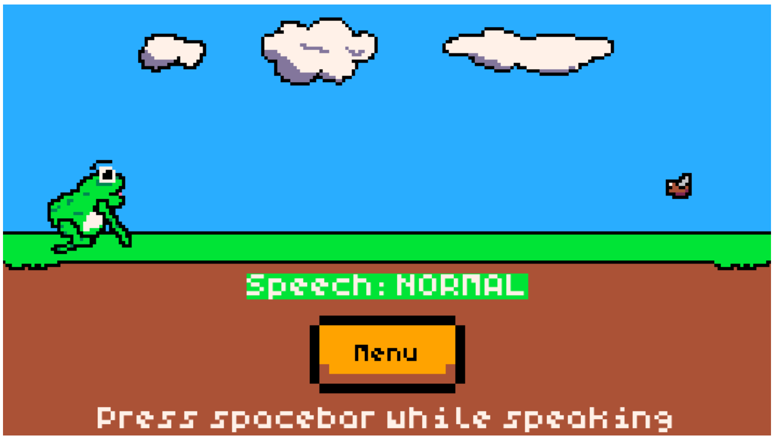


Figure 5.7: Increased difficulty, achieved by moving the fly away from the frog.

5.2.4. Implementation Details

5.2.4.1. Nasal Voice Detection

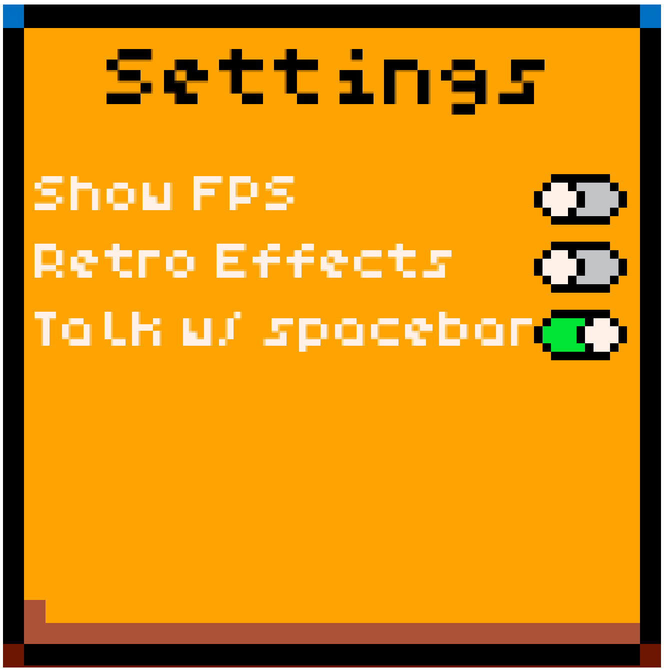
Because the models we trained are designed to make a decision from two classes (nasal and normal), the nasal voice detection classifies all the inputs either as a normal voice or a nasal voice. The problem is, the classification is done even when no one is speaking. To solve this problem, we added an option to talk while pressing the space key. This way, predictions are made only when the user speaks. We have also added a threshold for sounds below a certain loudness.

5.2.4.2. Processing Microphone Input

The microphone input library that we have used blocked the game until all the required audio is provided, so the game paused every time the microphone tried to get input. This problem was solved by letting the audio input and class predictions be done in a separate thread.

5.2.4.3. Graphical User Interface

The graphics are highly important in video games, especially when they are designed for children. We tried our best to make the game menu, settings etc. easy to understand. We created and used pixel-art images. The buttons, toggle switches, panels are all hand-crafted to make the GUI as consistent as possible.



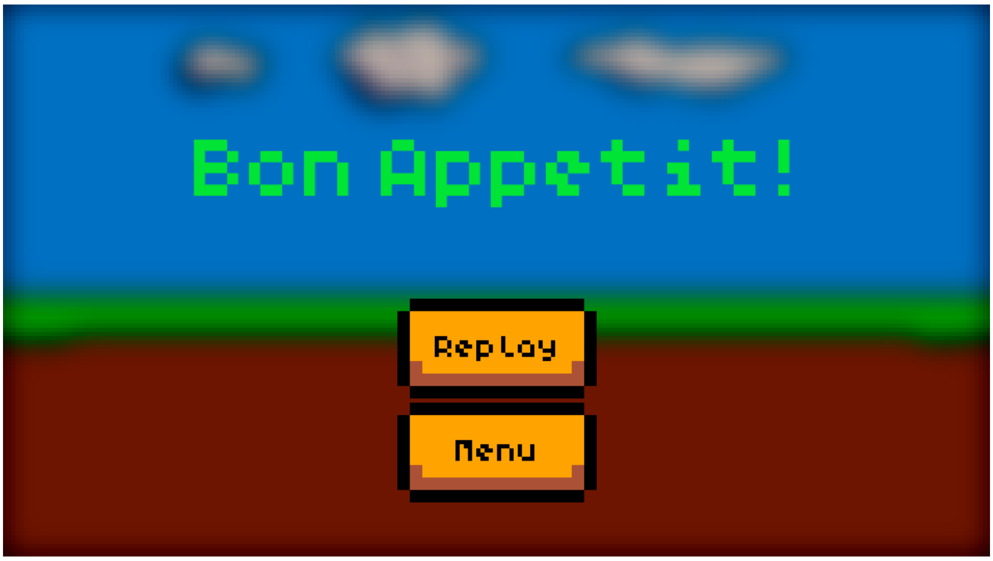
Settings

Show FPS ○

Retro Effects ○

Talk w/ spacebar ●

Figure 5.8: Settings menu.



Bon Appetit!

Figure 5.9: Ending of the frog game.

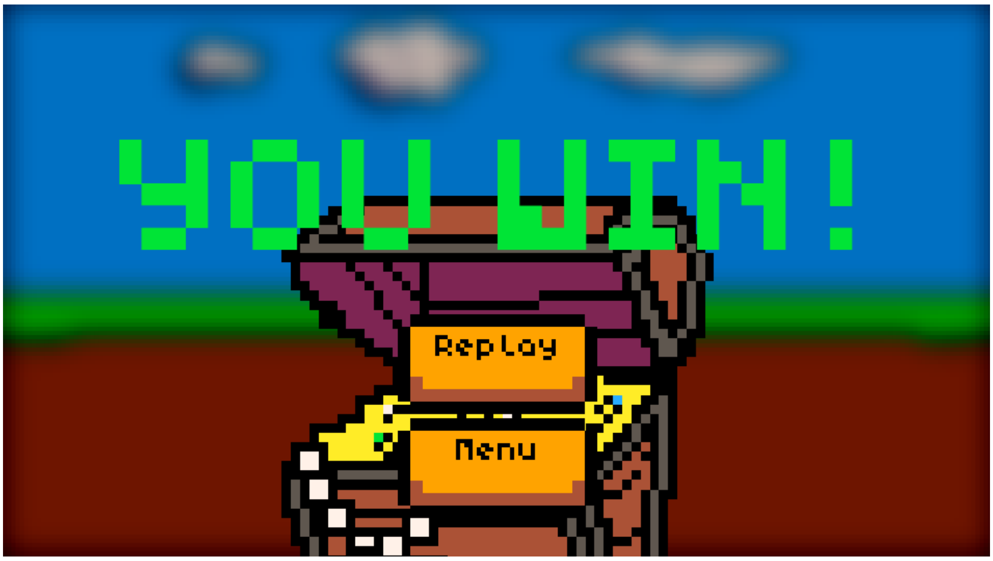


Figure 5.10: Ending of the diver game.

6. SUCCESS CRITERIA

For this project, we have set four criteria in total at the beginning of the semester. We have done our best to satisfy them in the period of time we had. The success criteria are as follows:

- At least 80% percent of the predictions should be accurate.
- A speed of 5 predictions per second should be reached.
- A dataset of at least 2 hours of audio with nasal and non-nasal speech should be created.
- At least two of three volunteers playing the game should give positive reviews.

The accuracy rates are mentioned in the Method section; we have reached high accuracy rates (97%) while testing with the data where the models have seen the person's voice, and lower accuracy rates (58%) with data where the models have never seen the voice of the person.

The model currently makes a prediction every 0.9 seconds, due to the input shape of the deep learning model. Our main focus was to enhance the accuracy rates, so the input of the models were not altered to increase the speed of the predictions.

The dataset we have is created using 11 different people's voices and the duration is above three hours. The game was played by the family members of the author, and has been remotely demonstrated to many people. The reviews of the observers were quite positive.

7. CONCLUSIONS

In this study we have aimed to develop a game that will help children with hypernasality to improve their normal speaking voices by visualizing when they are speaking normally and when nasally. We started collecting data, trained deep learning models and developed two mini-games to accomplish our goal.

We have clearly seen that the data collected is not enough for this audio classification problem. The models that are tested with people's voices that are familiar to them have scored high accuracy rates. However, they have failed to generalize the characteristics of the voices and predict the nasality of the other people.

Future studies may aim to use transfer learning, given that there is not have enough data. Another way to make the game work with variable voices might be by getting the users' speech data at the beginning of the game and training the model. This way the model will predict with high accuracy scores, making the game more playable.

BIBLIOGRAPHY

[1] Cagatay, M., P. Ege, *et al.*, “A serious game for speech disorder children therapy.,” 2012.

[2] Palanisamy, K. Singhania, D., Yao, and A., “Rethinking cnn models for audio classification.,” 2020.

[3] Sturm, B. L., Morvidone, M., Daudet, and L., “Musical instrument identification using multiscale mel-frequency cepstral coefficients.,” 2010.

[4] Dhanalakshmi, P., Palanivel, S., Ramalingam, and V., “Classification of audio signals using svm and rbfnn.,” 2009.

[5] Breebaart, J., McKinney, and M. F., “Features for audio classification.,” 2004.

[6] Grama, L., Rusu, and C., “Choosing an accurate number of mel frequency cepstral coefficients for audio classification purpose.,” 2017.

[7] Koolagudi, S. G., Rastogi, D., Rao, and K. S., “Identification of language using mel-frequency cepstral coefficients (mfcc).,” 2012.

[8] Sato, N., Obuchi, and Y., “Emotion recognition using mel-frequency cepstral coefficients.,” 2007.

[9] Vergin, R., O’Shaughnessy, D., Farhat, and A., “Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition.,” 1999.

[10] Costa, W., Cavaco, S., Marques, and N., “Deploying a speech therapy game using a deep neural network sibilant consonants classifier.,” 2021.

[11] Hasan, M. R., Jamil, M., Rahman, and M. G. R. M. S., “Speaker identification using mel frequency cepstral coefficients.,” 2004.

[12] Lopes, V., Magalhães, J., Cavaco, and S., “Sustained vowel game: A computer therapy game for children with dysphonia.,” 2019.

[13] Lan, T., Aryal, *et al.*, “Flappy voice: An interactive game for childhood apraxia of speech therapy.,” 2014.

[14] Ganzeboom, M., Yılmaz, *et al.*, “On the development of an asr-based multimedia game for speech therapy: Preliminary results.,” 2016.

[15] Shtern, M., Haworth, *et al.*, “A game system for speech rehabilitation.,” 2012.

[16] Rose Medical. “Nasality microphone.” (), [Online]. Available: `https://rose-medical.com//images/3D-nasometry-product-banner.jpg`.

[17] Dubey, A. Kumar, and et al., “Detection of hypernasality based on vowel space area.,” 2018.

[18] Akafi, Ehsan, M. Vali, and N. Moradi, “Detection of hypernasal speech in children with cleft palate.,” 2012.

[19] Wang, X., Yang, *et al.*, “Hypernasalitynet: Deep recurrent neural network for automatic hypernasality detection.,” 2019.

[20] Wang, X., Yang, *et al.*, “Automatic hypernasality detection in cleft palate speech using cnn.,” 2019.

[21] Smarty Ears Apps. “Apraxiaville.” (2017), [Online]. Available: `http://www.smartyearsapps.com/apraxia-ville/`.

[22] Little Bee Speech. “Articulation station.” (2018), [Online]. Available: `http://littlebeespeech.com/articulation_station.php`.

[23] Expressive Solutions. “Artikpix.” (2018), [Online]. Available: `http://expressive-solutions.com/artikpix/`.

[24] A. Parnandi, V. Karappa, T. Lan, *et al.*, “Development of a remote therapy tool for childhood apraxia of speech.,” 2015.

[25] Tiga Talk. “Tiga talk speech therapy games.” (2011), [Online]. Available: `http://tigatalk.com/app/`.

[26] Pocket SLP. “Pocket slp.” (2018), [Online]. Available: `http://pocketslp.com/`.

[27] Duval, J., Rubin, *et al.*, “Designing towards maximum motivation and engagement in an interactive speech therapy game.,” 2017.

[28] Librosa. “Mfcc spectrogram.” (), [Online]. Available: `https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html`.

[29] Muda, Lindasalwa, M. Begam, and I. Elamvazuthi, “Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques.,” 2010.

[30] Rose Medical. “Speech therapy games.” (), [Online]. Available: `https://www.rose-medical.com/speech-therapy-games.html`.