Classifying Hypernasality in Children with a Cleft Palate using a Convolutional Neural Network

Rebecca Svensson

Master's thesis
2022:E60

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Master's Theses in Mathematical Sciences 2022:E60
ISSN 1404-6342

LUTFMS-3456-2022

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/

Title
Classifying Hypernasality for Children with a Cleft Palate using a Convolutional Neural Network

Author
Rebecca Svensson

Figures
The displayed figures in this report are created by the author if nothing else is indicated

Acknowledgements

First and foremost, I would like to express my gratitude towards my supervisor Andreas Jakobsson, who has been guiding me throughout this project from start to finish. His weekly guidance, commitment towards the project and encouragement, made this thesis project possible. I would also like to earnestly acknowledge the sincere efforts and valuable time given by my co-supervisor Susanna Whitling in helping me understand speech difficulties related to cleft lip and palate. Magnus Becker, Måns Cornefjord, and Tofig Mamedov who are also involved or have been in the project should not be forgotten either, and I would like to express my gratitude towards them as well.

A special thanks is given to Anette Lohmander and Kristina Klintö for providing valuable data for the project, but also for providing guidance and knowledge very much needed.

I would also like to thank Liisi Raud Westberg for letting me use a few of her illustrations in this work.

Last but not least, I would like to thank my family and friends for their support. Without their endless patience and understanding I would have been lost.

Rebecca Svensson, Lund, June 2022

Abstract

In Sweden, about 150-200 children annually are born with some form of cleft lip and/or palate, making it the most common facial malformation in the country. Treatment often involves one or more surgeries and the speech development is followed up by a speech pathologist from the first year of life. One of the most common speech deviation for children with a cleft palate is hypernasality, which comes from the condition velopharyngeal insufficiency (VPI), which means that the soft palate cannot close the passage between the throat and nasal cavities properly. To classify the severity of VPI and hypernasality is hard, since listeners have different internal standards. This makes it important to develop an independent method to classify the severity.

This thesis studied two existing deep learning methods from another master's thesis on new data to see if it would be able to classify the severity of VPI, which is classified on a three point scale. The methods did not work well on the new data, but could be improved a bit by better data processing and some changes in the methods. The best performance for classifying VPI was a VGGish network which gave a file-wise accuracy of 57.11 %. A method to try and classify hypernasality was investigated as well. The best method found in order to classify hypernasality, which is classified on a four point scale, was to look at the vowels and to use mel spectrograms in a Convolutional Neural Network (CNN). Nasometry data was also given and together with the mel spectrograms an accuracy of 52.38 % was reached.

Sammanfattning

I Sverige föds ca. 150-200 barn årligen med någon form av läpp-, käk- och gomspalt (LKG-spalt), vilket gör LKG-spalt till den vanligaste ansiktsavvikelsen i landet. Behandling innebär ofta en eller flera operationer och talutvecklingen följs upp av logoped redan från det första levnadsåret. En vanligt förekommande talavvikelse för barn med LKG-spalt är hypernasalitet som kommer från tillståndet velofarynxinsufficiens (VPI), som innebär att den mjuka gommen inte kan sluta tätt mellan svalg och näshålor. Att klassificera allvarlighetsgraden av VPI och hypernasalitet är svårt, då lyssnare har olika interna standarder. Det gör det viktigt att utveckla en oberoende metod för att klassificera allvarlighetsgraden.

Detta examensarbete studerade två befintliga djupinlärningsmetoder från ett tidigare examensarbete på ny data, för att se om de kunde klassificera allvarlighetsgraden av VPI, som klassificeras på en tregradig skala. Metoderna fungerade inte bra på den nya datan, men de var möjliga att förbättra lite genom bättre databehandling och några förändringar i metoderna. Det bästa resultatet för att klassificera VPI, var ett VGGish-nätverk som gav 57.11 % noggrannhet. En metod för att försöka klassificera hypernasalitet undersöktes också. Den bästa metoden som togs fram för att klassificera hypernasalitet, som klassificeras på en fyrgradig skala, var att titta på vokalerna och att använda melspektrogram i en Convolutional Neural Network (CNN). Nasometridata fanns också tillgänglig och tillsammans med melspektrogramen uppnåddes en noggrannhet på 52.38 %.

Contents

Chapter 1

Background

1.1 Disposition

Background: In the background some background to cleft lip and palate, how voice and speech works and the theory behind deep learning are presented. What has previously been done in the area is also presented as well as the aim of the project and the project goals.

Methodology: The methods used in the thesis are presented in the methodology part of the thesis. The data that was used in the project is explained and how it has been processed in order to fit different deep learning methods.

Results: All the results achieved can be found in the result section.

Discussion and Conclusion: Under Discussion and Conclusion a discussion of the results are presented. Suggestions for further work can also be found.

1.2 Introduction

In Sweden, about 150-200 children annually are born with some form of cleft lip and/or palate (CLP), making it the most common facial malformation in the country [19]. Approximately 1 in 600 births in Sweden have CLP. CLP happens when the lip and/or hard palate and/or soft palate has not merged together properly. When the fetus is six weeks old, the hard palate and the lip is formed in the womb and between the eighth and tenth week the soft palate [4]. Hence, CLP can often be found on routine ultrasounds. A cleft

in the soft palate can however be harder to see, but is usually discovered directly after birth during the first control check. The reason for CLP is not known, but research indicates that it can be hereditary and that alcohol, medication, and smoking during the pregnancy can cause CLP [2].

There are many different kinds of CLP conditions and the severity can differ a lot. The lip, soft palate, and hard palate, can be affected in different combinations. A cleft in the lip and/or hard palate can be present on one side (unilateral) under the nose or on both sides (bilateral) [2]. Figure 1.1 illustrates unilateral and bilateral CLP were the lip, hard palate and soft palate are cleft.

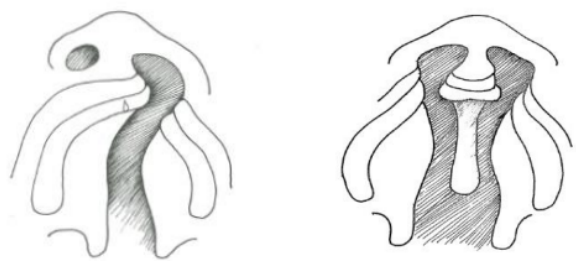

Figure 1.1: Illustration of unilateral and bilateral CLP. Illustration by Liisi Raud Westberg.

Treatment often involves one or more surgeries for cosmetic reasons and to help with eating, speaking and hearing. Even with surgery, speech deviations occur and the speech development is followed up by a speech pathologist from the first year of life [2][4]. One of the most common speech deviations for children with a cleft palate is hypernasality. It comes from the condition velopharyngeal insufficiency (VPI), which means that the soft palate cannot close the passage between the throat and nasal cavities, making the nasal cavity a resonator and the sound nasal [4]. Figure 1.2 illustrates the insufficient closure of the soft palate when air is able to pass from the throat to the nasal cavities.

In order to evaluate the children's speech, speech pathologists use different tests as for example SVANTE (SVenskt Artikulations och Nasalitets-TEst). The child is recorded while saying words by naming pictures, repeating standard sentences or retelling a story. The speech pathologist is listening to the recording in order to find deviations in the speech. To classify the severity of VPI, and especially hypernasality auditively is hard. When rating voice quality, listeners have different internal standards. However, experienced clinicians may develop their own standard references and they have hence

better reliability than inexperienced, but rating hypernasality may still be hard [9].This makes it important to develop an independent method to classify the severity of hypernasality.

In order to explore deep learning methods to cleft palate speech a master's thesis was done in 2021 in order to classify VPI by using a Convolutional Neural Network, a BiLSTM network and a T.L VGGish network [5]. This master's thesis is a continuation of the same project, but with new data.

Figure 1.2: The air passing up to the nasal cavity when the soft palate does not close properly. Illustration by Liisi Raud Westberg.

1.3 Aim and Project Goals

The overall aim of the project was to create an unbiased deep learning method to classify velopharyngeal insufficiency and hypernasality in children with a cleft palate. Since it is hard to classify hypernasality auditively, the project was done in order to help speech pathologists and other medical staff to be able to give the best treatment possible to children with a cleft lip and palate.

The project goals were to:

- Investigate how the Convolutional Neural Network implemented in 2021 works on new data [5].
- Depending on how good the Convolutional Neural Network works on the new data, to try and improve the performance.
- Implement a model that is able to classify the amount of hypernasality present in the children's speech.

1.4 Voice and Speech

In order to understand how CLP speech works and how it differs from normal speech, a short presentation of how the speech system works is presented here. In figure 1.3 the organs that is present during normal speech production are illustrated. Voice is under normal circumstances produced when airflow coming from the lungs passes through the vocal folds and the vocal tract. The vocal folds vibrate in order to convert the air to acoustic energy, which are acoustically filtered by the vocal tract. In order to produce different sounds, the way the vocal folds vibrate and how the vocal tract looks above the larynx has to be changed. For example by moving the tongue, soft palate, jaw and lips [11]. It is here the condition velopharyngeal insufficiency comes in since it makes it difficult for the child to close the opening between the oral and nasal cavities. This leads to increased flow of air through the nose instead of the mouth, which can cause hypernasal speech [4].

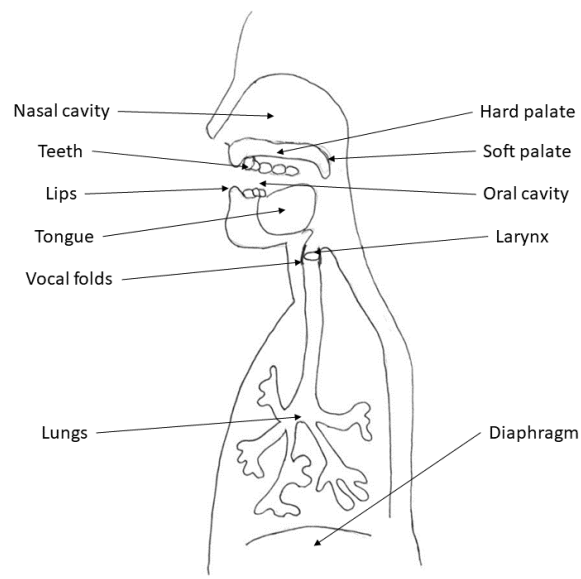

Figure 1.3: Overview of vocal tract. Inspiration to sketch from [11].

1.4.1 Hypernasal Speech

In hypernasal speech due to velopharyngeal insufficiency, vowels are affected. In all languages vowels are used. However, the occurrence of vowels differ a lot between languages which might influence the assesment of cleft palate speech

since hypernasality is mainly perceived on vowels. The distribution of the different vowel types (high/low/front/back) also varies between languages. In cross linguistic studies of cleft palate speech the height of the vowels plays a big role since studies have shown that hypernasality is easier to distinguish in high vowels such as [i] and [u], than in low vowels such as [a] or [α]. In cross linguistic studies to assess hypernasality it is important to have speech samples with comparable phonetic contexts and similar vowel qualities [9].

Harmonic ratio is a measure between 0 and 1 of the amount of energy in the tonal part of the signal compared to the the amount of energy in the total signal [15]. Since vowels typically contain more energy than consonants harmonic ratio can be used to pick out vowels and high energy sounds, which is usually more affected by hypernasality.

1.4.2 Time-frequency Representation and Mel Spectrogram

An audio signal is measured by how air pressure varies over time (the time-domain), giving a waveform for the signal. The signal is a combination of several single frequency waves, adding up to the resulting amplitudes [18]. In figure 1.4, the wave form of the phrase "*Titti tittar på TV*" can be seen, when uttered by a 5-year-old. The sampling rate is how many samples that is measured per second. Different sampling rates can be used when measuring audio, for example 16kHz, 44.1kHz, or 48kHz. However, in order to reliably restore the signal the highest frequency that can be restored is the Nyquist frequency, which is half of the sampling frequency [20]. For example, when using a sampling rate of 16kHz, frequencies up to 8kHz will be reliable for investigation.

An audio signal can be transformed from the time-domain to the frequency-domain, in order to visualize the frequencies present. This is done by using the fast Fourier transform which decompose the signal into sine and cosine signals in order to find the frequencies that compose the signal and their amplitudes [18]. This way of representing a signal is called a spectrum. However, an audio signal may only be considered stationary for a short period of time, usually about 10-30 ms [21], and hence the frequency content varies over time. In order to visualize this several spectra are computed for small overlapping window segments of the audio signal, using the so-called Short-time Fourier transform (STFT). The result is called a spectrogram [18]. The spectrogram of "*Titti tittar på TV*" can be seen in figure 1.5. The spectrogram is a visualization of how the amplitude for different frequencies varies over time. As can be seen in the figure, the frequency-axis is shown in the log scale and the amplitude in decibel. The amplitude is the third dimension

and is represented by colour.

The human ear does not perceive frequencies in a linear way. Differences in lower frequencies are easier for humans to detect than differences in higher frequencies. The difference between 500Hz and 1000Hz is easily distinguished, but not the difference between 10000Hz and 10500Hz. The mel scale was created in order to compensate for this, making the distance in pitch sound equally distant to the listener. If the frequencies are converted to the mel scale for a spectrogram, a mel spectrogram is created [18]. In figure 1.6 the mel spectrogram of the sentence "*Titti tittar på TV*" is shown.

1.5 Deep Learning

Deep learning is a machine learning technique that has been given a lot of attention lately due to its ability to solve classification tasks with very good accuracy. Deep learning use essentially a human approach, to learn by example in order to for example distinguish an image of a dog from an image of a cat. The models learns to classify sound, images or text by training on a large set of labeled data. In order to learn, the models use many (deep) layers of neural network architectures, and hence no manual feature extraction is needed [14]. In figure 1.7, an overview of how a neural network is structured can be seen. The layers consist of nodes that are connected. The inputs could be images for example and the output the classifications of the images.

1.5.1 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a type of a deep neural network, which is suitable for, for instance, 2D data such as images, since it typically uses 2D convolutional layers. The network extracts features during training on a training set of images, and no features are pretrained [14].

In order to get a CNN to work well on a data set, different types of layers need to be used in an order suitable for the task. The first layer of a CNN is often the image input layer which reads the images. The images are typically assumed to have the same size when they enter the network. The last two layers are typically a softmax layer and a classification layer which returns the classification for the image [13]. Other important layers are explained in more detail below.

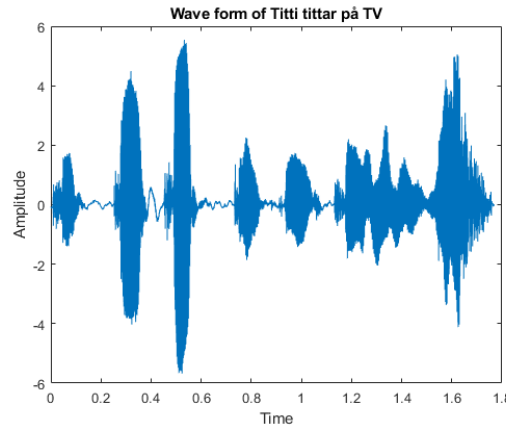

Figure 1.4: The wave form of the sentence "*Titti tittar på TV*".

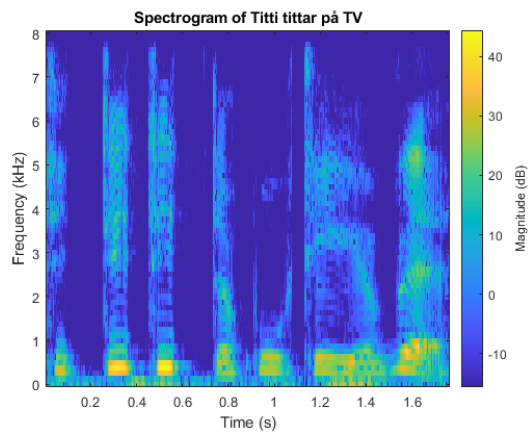

Figure 1.5: The spectrogram of the sentence "*Titti tittar på TV*".

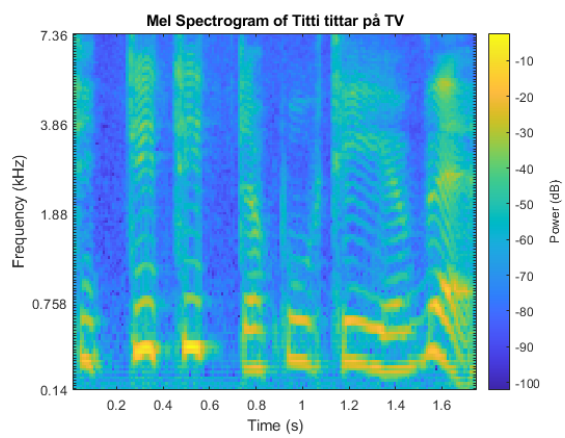

Figure 1.6: The mel spectrogram of the sentence "*Titti tittar på TV*".

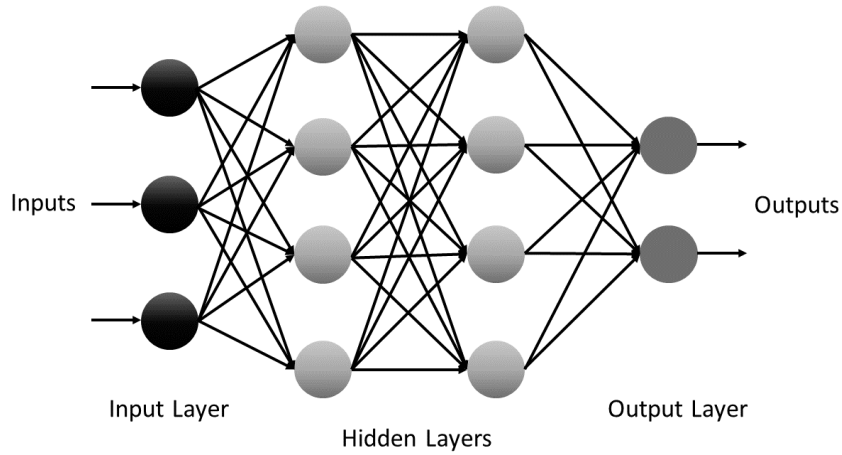

Figure 1.7: An overview of a neural network.

The Convolutional Layer Depending on the data the number of convolutional layers varies. One or more layers can be used. The input to a convolutional layer is the input images to the network or the outputs of the previous layer. The neurons of the convolutional layer connect to subregions in the input. A filter, a set of weights, of a pre-defined size and step size slides over the input both vertically and horizontally and repeats the same computation in all regions. The dot product of the weights and the input is computed and a bias term is added. Depending on the filter size and the stride size, the regions the neurons connect to may or may not overlap. The layer learns in this way the features present in the regions.

A filter contain $h \cdot w \cdot c$ weights, where h and w are the height respectively width of the filter and c the number of channels. In a colour image the number of channels is 3 and for a black and white image there is only one channel. As the filter moves across the input the same weights and bias are used for the convolution, which creates a feature map. The number of feature maps is equal to the number of filters and is the result of a convolution using different weights and bias. The number of parameters is hence $(h \cdot w \cdot c + 1) \cdot NbrF$, where $NbrF$ is the number of filters, with the one accounting for the bias.

For a convolutional layer, the output height and width can be computed as

$$\frac{InputSize - FilterSize + 2 \cdot Padding}{Stride} + 1 \tag{1.1}$$

Padding is the amount of padding applied to the borders vertically and horizontally. Padding means that values are appended to the borders of an image in order to increase the size. The output size can be controlled depending on how much padding is applied.

The number of neurons in a feature map is the product of the output height and width. The total number of neurons in a convolutional layer is the map size times the number of filters [13].

The Batch Normalization Layer The batch normalization layer is usually used between convolutional layers and ReLU layers. The layer normalizes a mini-batch of data by subtracting the mean of the mini-batch and dividing by the standard deviation of the mini-batch. The layer then shifts the input an offset β and scales it by a scale factor γ. Both β and γ are learnable parameters that updates during the training of the network [13].

The ReLU Layer A ReLU layer performs a threshold operation on the input. The threshold is zero and values less than zero is set to zero. Usually the ReLU layer follows a convolutional layer and a batch normalization layer [13].

The Max Pooling Layer Max pooling is a type of pooling layer which usually follows a convolutional layer in order to down-sample. The down-sampling reduce the number of parameters to be learned in the next layer, and it also help with overfitting. A max pooling layer takes the maximum value of the input in a rectangular region which moves across the input. The size of the rectangular region is chosen when the layers are defined [13].

The Fully Connected Layer

The fully connected layer/s usually follow the convolutional layers. The input is multiplied with a weight matrix and then a bias vector is added. The fully connected layer combines all the features that the previous layers have learned by letting all neurons in the previous layer connect to all neurons in the fully connected layer. The last fully connected layer combines all the features in order to classify the image and hence the last fully connected layer of a network have the same output as the number of classes present in the data set [13].

1.5.2 The Pre-trained Network VGGish

VGGish is a pre-trained network trained by Google on 70 million YouTube videos in order to classify audio [8]. The YouTube audio was first resampled

to 16kHz and to mono. A spectogram was then computed by using the Short-time Fourier transform. A periodic Hann window was used with a window size of 25 ms and a hop size of 10 ms. A mel spectrogram was then computed by filtering the spectogram through a mel filterbank of 64 mel bins in the frequency range of 125-7500 Hz. A log mel spectrogram was then computed. In order to avoid taking log of zero an offset of 0.01 was added to the mel spectrum before the natural logarithm was taken of the mel spectrogram. The log mel spectrogram was then cut into frames of 0.96 seconds, with 64 mel bands and 96 frames of 10 ms each. The log mel spectrograms were then put into a network with 24 layers. Six of the layers were convolutional layers and three were fully connected. In between these, there are max pooling layers and ReLU layers [16].

1.6 Literature Review

Using deep learning in order to classify cleft palate speech is a topic explored by different research groups. The fact that velopharyngeal insufficiency and especially hypernasality is hard to classify auditively [9], makes the topic very attractive to explore.

One paper that explores the use of spectrograms and Convolutional Neural Networks in order to classify hypernasality is the paper *Automatic Hypernasality Detection in Cleft Palate Speech Using CNN* written by a Chinese research group [21]. The method they use is to first segment out the vowels from the audio data set, since these are affected by hypernasality. The vowels were then used to create spectrograms of 20 ms and with a 10 ms overlap. The spectrograms were then used in a Convolutional Neural Network in order to classify if the frame should be deemed hypernasal or not. They managed to get good results on both a data set spoken by children and a data set spoken by adults. For the children, the average F1-score was calculated to be 0.9485 when using cross-validation with the method of leaving one patient out for validation while training. The F1-score was calculated as

$$\text{F1} - \text{score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{1.2}$$

where the precision was defined as TP/(TP+FP) and recall as TP/(TP+FN). TP stands for True Positive and refers to the probability of normal speech being classified as normal speech. FP stands for False Positive and refers to the probability of normal speech being classified as hypernasality. FN

stands for False Negative and refers to the probabilty of hypernasality being classified as normal speech [21].

A master's thesis project at the medical program done in Edinburgh called the CLeft-Associated Resonance Imitation Study (CLARIS report), also strengthen the method of using vowels [17]. The report use speech samples from experienced speech pathologists imitating hypernasal speech. Different vowels are examined for Swedish, English and Portuguese using mel spectrogram. It is possible to see a difference in the mel spectrograms and good results are obtained when using different deep learning algorithms, one of them being CNN.

In 2021, a master's thesis titled *Exploring Deep Learning Approaches to Cleft Lip and Palate Speech* aimed to classify velopharyngeal insufficiency, using speech data from 5 and 10 year old children [5]. First, the children's voice had to be segmented out since the audio contained a lot of other people speaking as well. The speech data left was divided into segments of 0.2 s and only frames with high energy were kept. The frames kept were then transformed into mel spectrograms and used in a Convolutional Neural Network. The pre-trained networks VGGish and BiLSTM were also tried. The CNN performed best with a file-wise accuracy of 89.76 %. Regrettably, this result has, as part of this work, been discovered to be erroneous. For the two pre-trained networks, VGGish had a file wise accuracy of about 57.67 % and BiLSTM 56.05 %. When reexamining the architectures for the pre-trained networks, they were found to be legit.

Other deep learning methods to classify hypernasality can also be found in litterature. One of them can be found in the paper *A Deep Learning Algorithm for Objective Assessment of Hypernasality in Children With Cleft Palate* [12]. The method uses an algorithm that automatically measures hypernasality in speech. A deep neural network on healthy speech was trained to find nasal acoustic cues and then tested on a data set with children with a cleft palate. The results showed that the measured hypernasality significally correlated with the perceptual ratings.

Chapter 2

Methodology

2.1 Software

The software used in this project was Matlab 2022a with simulink and the deep learning toolbox. The free recording and editing software Audacity(R) version 3.0.0 was used for editing audio files [1]. The free speaker diarization Python toolkit pyAnnote was also used [6].

2.2 The Data Sets

The data used in this project was provided by Professor Anette Lohmander at Karolinska Institutet in Stockholm. The data comes from two different studies; the Scandcleft project (SC) [3] and the Intercenter material (IC) [10]. The Scandcleft project was a multicenter study that started in 1997 between centers in Aarhus and Copenhagen in Denmark, Oslo and Bergen in Norway, Gothenburg, Linköping and Stockholm in Sweden, Helsinki in Finland, and Belfast and Manchester in the United Kingdom.

The Intercenter material comes from a Swedish study with children born 2008-2010 from the six cleft lip and palate centers (Gotherburg, Linköping, Malmö, Stockholm, Umeå, Uppsala-Örebro) in Sweden. All children were 5 years old when the voices were recorded, had unilateral cleft lip and palate, and had had some sort of correction surgery.

The children in the Scandcleft project were both audio taped and video recorded. One speech pathologist sat opposite the child and another recorded

the child from behind the first speech pathologist, making the camera facing the child. Two microphones were placed about 40 cm from the table edge, one in front of the child and one on the side of the camera. The audio recordings were recorded using a DAT (Digital Audio Tape) tape recorder or one of comparable quality with a condenser microphone [3]. In the IC study, the children were recorded in a quite room at one of the Hospitals participating, using a condenser microphone [10]. Since the IC-study were done a couple of years after the SC project, the recordings are typically of better quality. In the SC-study, the sampling frequency was 44100 Hz for the data distributed to be used for testing hypernasality and 48000 Hz for velopharyngeal insufficiency. All of the recordings in the SC-project were recorded in stereo. In the IC-study, the sampling frequency varies between 16000-48000 Hz with some recordings being in stereo and some in mono. In this work, all recordings were re-sampled to 16000 Hz and to mono.

2.2.1 Data for Velopharyngeal Insufficiency

In both the SC-project and the IC-study, recordings of the so-called bus story were used to evaluate the velopharyngeal insufficiency (VPI). The children repeated a story about a bus with some help from a speech pathologist. In the IC-study, recordings from the Swedish articulation and nasality test (SVANTE) were also used in order to evaluate the velopharyngeal function. SVANTE consists both of single words that the children are naming by looking at pictures and by repeating sentences after a speech pathologist. Here, 13 sentences were used and each of them contained a recurring consonant. Data from the sentence repetition task was used to evaluate VPI [10].

The velopharyngeal function was graded on a scale 1 to 3, with 1 = Competent, 2 = Marginally incompetent and 3 = Incompetent. Competent, in this case, means that the velopharyngeal function of closing the passage between the nasal cavities and the oral cavities is satisfactory. Marginally incompetent means that the function is not completely satisfactory and incompetent that it is unsatisfactory. In the evaluation the speech pathologist took speech symptoms such as hypernasality, nasal air leakage, and weak articulation into account. In the IC-study, the evaluations were done by five speech pathologists from five different cleft centers in Sweden and the median values were used as the final ratings. In the Scandcleft project, three speech therapists evaluated the children and the median value was used as well.

All children in the IC-study were present in both the hypernasality ratings and the VPI ratings, with 57 children participating. However, three children repeated another story than the so-called bus story and for one child only

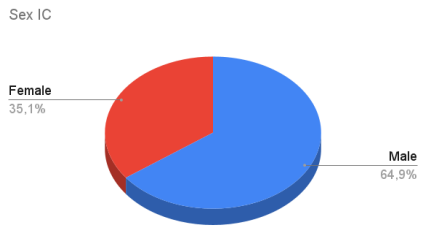

Figure 2.1: The distribution of gender for the IC data.

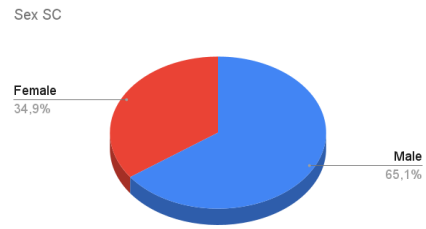

Figure 2.2: The distribution of gender for the SC data.

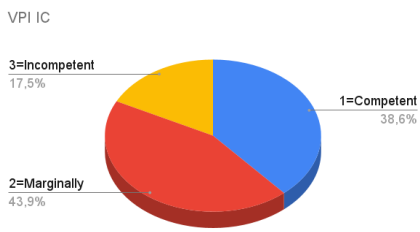

Figure 2.3: The VPI distribution for the IC data.

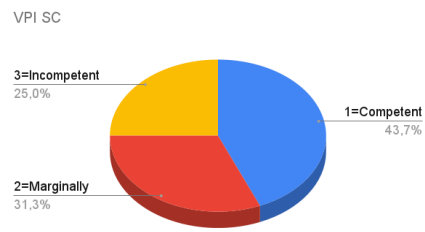

Figure 2.4: The VPI distribution for the SC data.

the sentence repetition were provided. The gender distribution can be seen in figure 2.1 and the VPI distribution can be seen in figure 2.3.

Some of the children in the SC-project were present in both the hypernasality ratings and the VPI ratings, but some were only present in one of the data sets. The gender of the children are given in figure 2.2. For the VPI ratings, 342 audio recordings were provided. The rating distribution is shown in figure 2.4. As can be seen, both the gender distribution and the VPI distribution are similar for both data sets.

As mentioned, the median value from the ratings done by the speech pathologists for each child are used as being the true class for that child. However, the individual ratings done by the speech pathologists may differ from this value. In order to examine how well the speech pathologists relate to the median value a confusion matrix may be used. For the IC-study it can be seen in figure 2.5. In the SC-study the individual ratings for each child was not present and hence no such confusion matrix could be presented. In the figure, "True Class" relates to the median values and "Predicted Class" to the ratings done by the speech pathologists. Overall the individual ratings coincides about 80 % of the times for the VPI ratings for the IC-study.

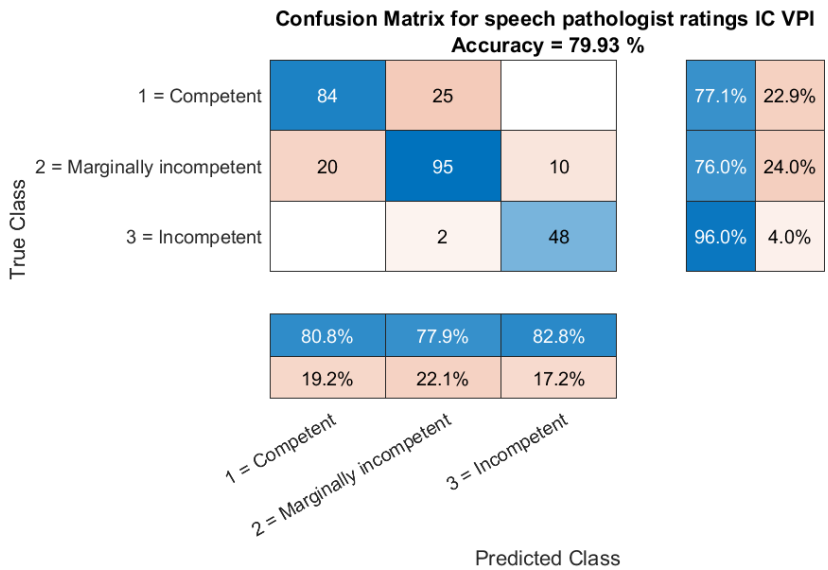

Figure 2.5: The distribution of ratings done by the speech pathologists for VPI, in the IC-study.

2.2.2 Data for Hypernasality

In the Scandcleft project, the data used to evaluate hypernasality was a 9 word sequence spoken by each child. The 9 words used for the different languages can be found in table 2.1. The different words were selected in order to detect hypernasality, and had to be present in the children's vocabulary. In order to get the children to say the words, pictures were presented to the child and the target word were elicited by naming. If the child did not know the word, semantic prompting was used. If that did not work either, the child was asked to repeat the word after the speech pathologist. However, since the children were only 5 years old, not all children are saying all the words.

Hypernasality was evaluated according to a four-point scale with 0=Normal, 1=Mild, 2=Moderate, and 3=Severe. The evaluation was done by three speech pathologists for each child in the Scandcleft project. One speech pathologist evaluated all children independent of language. The other two had the same mother tongue as the child. The median values of the three ratings were used in the final rating. The distribution of the ratings can be seen in figure 2.7. The gender of all the children in the SC-data for hypernasality were not given, but from the ones known it could be concluded that around 60 % were male and 40 % female. The total number of audio

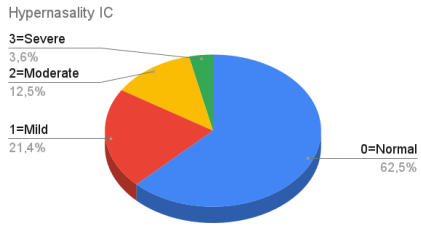

Figure 2.6: The distribution of the hypernasality ratings for the IC data.

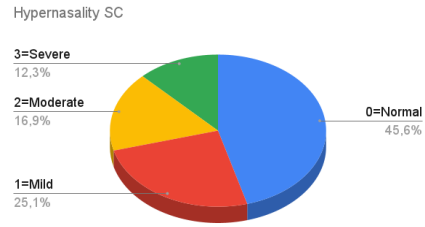

Figure 2.7: The distribution of the hypernasality ratings for the SC data.

Table 2.1: The 9 words used in the different languages in the Scandcleft project to evaluate hypernasality.

Swedish	**Norwegian**	**Danish**	**Finnish**	**English**
hus	biler	is	pallo	pea(s)
filar	is	fugle	pilli	peep
pippi	fire	piger	puu	bee
ljus	dyr	hus	pää	tea
bilar	lys	biler	paivä	teeth
teve	kuer	puder	peili	key
gul	gul	gule(rödder)	talo	geese
vit	hus	lys	talvi	feet
duva	fugler	gul	taulu	knee

files in the SC-data for hypernasality, were 390 files.

In the Intercenter study, the Swedish articulation and nasality test SVANTE was used [10]. Data from the sentence repetition task was used in this project and the hypernasality evaluations had also been done on the sentences. The same four-point scale that was used in the Scandcleft project were used in the Intercenter study, with 0=Normal, 1=Mild, 2=Moderate, and 3=Severe. The evaluation was done by five speech pathologists from five different cleft centers in Sweden, with the median values being used as the final rating. The distribution of the ratings can be seen in figure 2.6. The gender distribution for the SC and IC data are almost the same. In total, there were 57 audio files from the IC study.

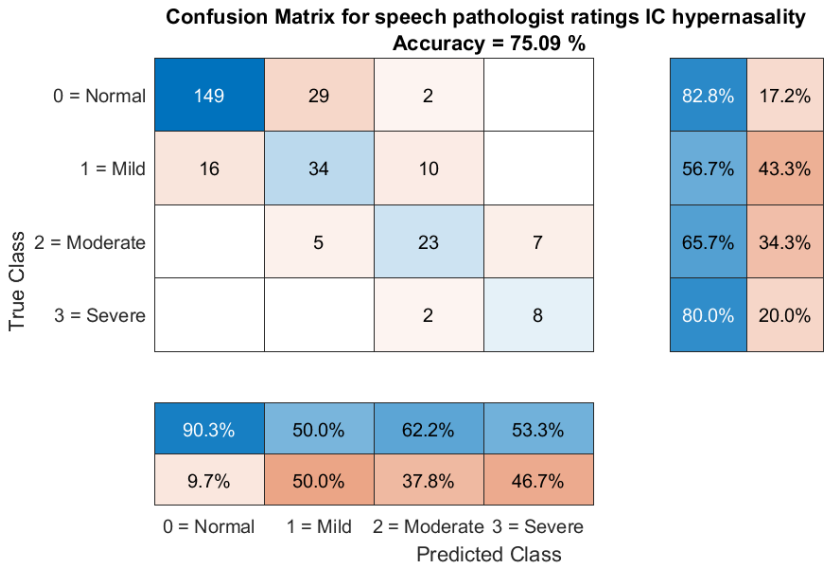

Figure 2.8: The distribution of ratings done by the speech pathologists for hypernasality, in the IC-study.

As for the VPI ratings, the median value from the ratings done by the speech pathologists for each child are used as being the true class for that child. The individual ratings for hypernasality done by the speech pathologists may differ from the median value. In order to examine how well the speech pathologists relate to the median value a confusion matrix may be used. For the IC-study it can be seen in figure 2.8 and in the SC-study it can be seen in figure 2.9. In the figures, "True Class" relates to the median values and "Predicted Class" to the ratings done by the individual speech pathologists. Overall, the individual ratings coincides with the median value about 75 % of the time for the IC-study and about 72 % for the SC-project. However, in the IC-study if the rating does not coincide with the median, the rating is in most cases only one step of, while for the SC-project, ratings may differ from the median value as much at 3 steps in some cases.

For some of the children in both the Scandcleft project and in the Intercenter study, nasometry data exists. Nasometry is an acoustic computer-based method for quantifying how much sound is emitted through the nose contra the mouth. The measure for nasometry is called nasalance and is calculated as

$$\%Nasalance = \frac{Nasalenergy}{Nasalenergy + Oralenergy} \cdot 100 \tag{2.1}$$

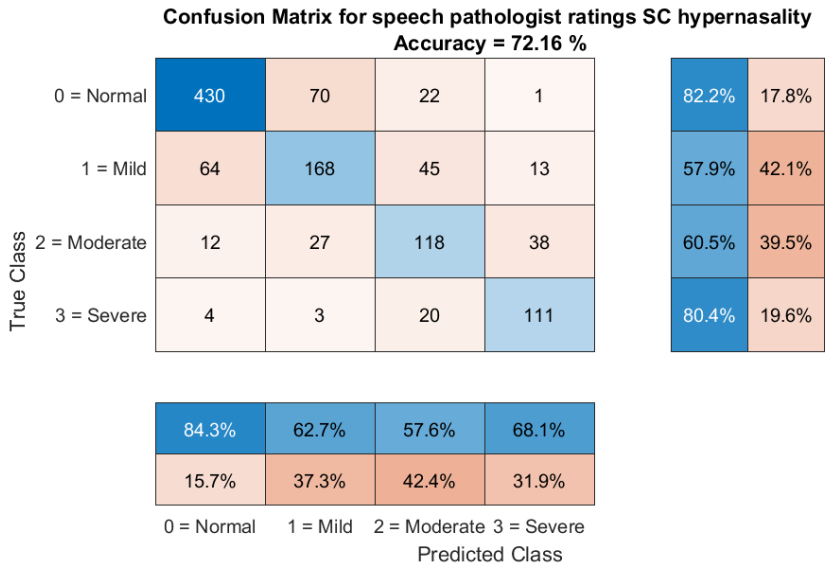

Figure 2.9: The distribution of ratings done by the speech pathologists for hypernasality, in the SC-study.

Nasalance is hence the percentage of the speech signal energy coming out of the nose divided by the total energy from both the nose and mouth. A device called nasometer is used for measuring nasalance. The nasometer is placed on the head of the patient and a plate separates the nose and the mouth and the oral and nasal speech signals are recorded by separate microphones. The patient is then saying words with high pressure vowels [7]. In the Scandcleft project a mean over the nasalance value for saying the 9 words five times were used [3]. A similar method were used in the Intercenter study. Typically, 21 % can be used as a breakpoint between normal speech and hypernasal speech for Swedish [7]. This value were used for all languages since there are no big differences between the languages used in the Scandcleft project. The distribution of nasalance scores for the different ratings (0-3) were plotted for the Scandcleft project in figure 2.10 and for the Intercenter study in figure 2.11. In the IC-study, a clear separation can be seen at the breakpoint between normal speech and hypernasal speech for ratings 1-3, all of which have a nasalance score above the breakpoint. It can also be seen that hypernasality scores 0 and 1 typically have a nasalance value below 40 % and scores 2 and 3 over 40 %. This trend can sort of be seen in the Scandcleft project as well, but not as clear.

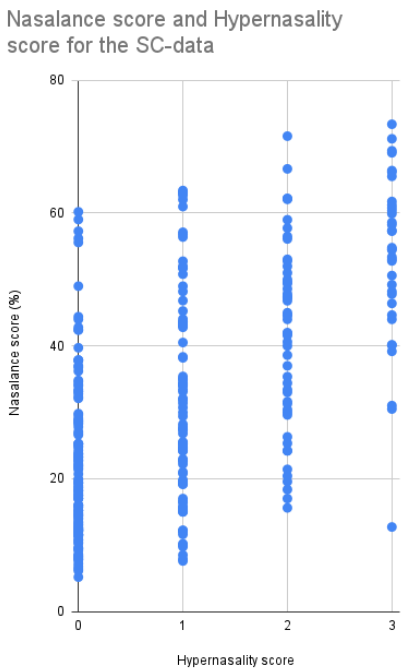

Figure 2.10: Nasalance score and Hypernasality score for the SC-data.

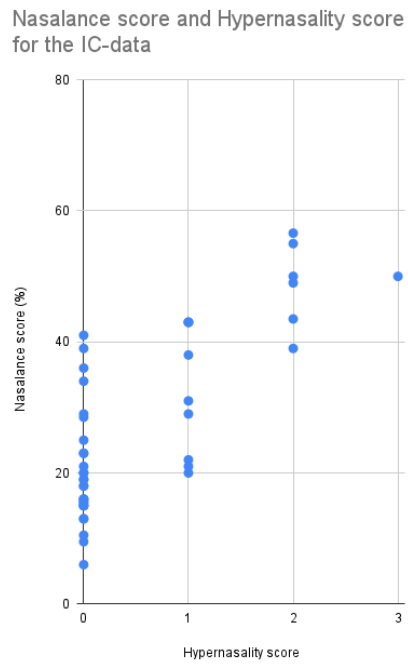

Figure 2.11: Nasalance score and Hypernasality score for the IC-data.

2.3 Implementation of Methods

In order to try and fulfill the project goals, different tests were set up. First the Convolutional Neural Network from the master thesis done in 2021 was implemented according to the specifications in [5]. The VGGish method was also evaluated since there seemed to be some issues with the CNN implemented 2021. It was also tried to improve these methods to receive better results. Finally, hypernasality was tested. Here, both a binary data set and a test set with all 4 categories were used.

2.3.1 Earlier Method for Velopharyngeal Insufficiency

In the method introduced in [5], each recording should be re-sampled to a 16 kHz sampling rate and the amplitude normalized between -1 and 1 in order to even out differences in microphone placement. In our implementation, the re-sampling took place first in a later stage since all files had the same sampling rate already (48000 Hz) and the sampling rate would not affect the initial segmentation of the audio. A common problem in all audio recordings used in the project were that not only the children are speaking. The speech pathologist and sometimes parents, siblings and others can also be heard. These voices had to be removed in order to only have the children's speech present in the recording. The removing was done by using a python toolkit called pyAnnote for speaker diarization [6], partitioning the audio file into segments according to speaker identity [5]. The output was a list of the times when someone is speaking, as well as who is speaking. The program only identify the number of speakers and labels them speaker A, speaker B etc. and cannot distinguish between child, parent, sibling and speech pathologist. This method was tried on the data. New files with only the children and a few utterances of other people were created with the diarization and by listening for which speaker was the child. A simple pitch averaging to see which speaker was the child was not done as in [5], since all files had to be double-checked anyways. For the IC-data, the method did not work at all and none of the files were successfully segmented as only one speaker was identified when the files contained two or more voices. For the SC-data, about half of the data was deemed segmented enough to be used. PyAnnote had been updated since it was used in [5] and the updated program was used since it gave better segmentation than the earlier version.

After the children's speech had been segmented out the files were re-sampled to 16 kHz. Each file was then divided into frames with size 0.2 s. Segments with a harmonic ratio above 75% were kept and segments below were thrown

away. A new file was created with the segments [5].

The final files were sent to the Convolutional Neural Network with the same layers, image size and training options as described in [5]. The mel spectrograms operated on approximately 0.2 s frames, that had the size 96x64 going into the network. Cross-validation with 10 iterations were used to validate the result with 70% of the data used for training and 30% for validation.

In order to use the pre-trained network VGGish, the audio had to be processed in the same way as the audio the network was trained on, i.e. log mel spectrograms of size 64x96. The layers also had to be modified slightly in order to classify the output into three classes for VPI. Therefore the last layer was removed and a fully connected layer and a softmax layer was put at the end. The same preprocessed audio files with a high harmonic ratio as for the CNN were used.

2.3.2 Updates made for Velopharyngeal Insufficiency

As pyAnnote did not work at all for the IC-data and since only half of the recordings in SC could be used, it was concluded that the method of speaker diarization by using pyAnnote was not good enough for the data sets in this project. It was noted that pyAnnote can be edited to fit specific problems better, but it was not attempted. The removal was instead done manually by using Audacity [1] to get as good files as possible. When each file was manually looked at, it was concluded that the quality for some of the files were too bad to use at all, due to them being saturated, noisy or other people speaking in the background. These files were removed from the data set and only files deemed to be good enough after the segmentation were kept. These files were sampled to mono, re-sampled to 16000 Hz and normalized to peak amplitude -1.0dB. Silence were also removed, in order to have as information dense files as possible.

The remaining data set contained 308 files, with roughly the same distribution between the classes as before. The final data set contained both the files from the IC-study and the SC-project put together. It should be noted that since the VPI was rated on both the bus story and the sentences in the IC-study the IC-files contained both. The files from the SC-study only contained the bus story. Only one file was removed from the IC-study, and the rest of the insufficient files were removed from the SC-project.

Both CNN and VGGish were tried. The final networks and parameters can be found in the appendix and are similar to the ones used in [5]. In the

CNN the same method as before with only using frames with a harmonic ratio over 75 % were used, but instead of looking at 0.2 s frames, smaller segments were used to get a more refined data set. The frames were 20 ms and only frames with a harmonic ratio over 75% were kept and used to create new data files. The files were then used in the CNN, but with 0.2 s frames and with an overlap of 50 %. In the VGGish network the whole files were used. VGGish is constructed to read raw data and the method worked better that way. An overlap of 75 % between frames were used. The classes in the data set were not equally distributed, which may influence the training of the network. Hence instead of using 70% of each class for training and the rest for validation, the number of files in the smallest class decided how many files were used for training and validation. The class with the least number of files were 3 = Incompetent with 73 files, of these 60 files were used for training and 13 for validation, which is roughly 80 % for training. For the other classes 60 files were also used for training and the rest for validation. As before cross-validation with 10 iterations were used in order to validate the result.

2.3.3 Method for Hypernasality

The audio files used for hypernasality were the 9 words from the SC data and the sentences from the IC data. The 9 word recordings had already been edited before this project and the words had been segmented out when possible. Some of the recordings were however saturated and noisy and some of the audio files were too bad to use and were hence discarded. The rest of the files from the SC-study and the IC-study were manually segmented with Audacity [1] to only contain the children's speech. They were also sampled to mono, re-sampled to 16000 Hz and normalized to peak amplitude -1.0dB. Even if the data sets from the two studies looked very different from each other the grading scale of hypernasality was the same and hence the data sets were combined in order to create a bigger data set. From this bigger data set four data sets were constructed as follows:

1. All files with ratings from 0-3.

2. All files which have nasometry data and their nasometry data, with ratings from 0-3.

3. Only files rated 0, 2 or 3, which were given the new ratings 0=Normal (the ones rated 0) and 1=Hypernasal (the ones rated 2 or 3) .

4. Only files which have nasometry data and were rated 0, 2 or 3 were used. The new rating 0=Normal (the ones rated 0) and 1=Hypernasal (the ones rated 2 or 3).

In order to retrieve the important parts of the speech to classify hypernasality, the vowels had to be retrieved since hypernasality typically shows up in vowels according to previous studies. The vowels were retrieved by using harmonic ratio with a threshold of 90 %. The threshold was chosen by listening, and 90 % worked well for most files. Segments of 25 ms with a harmonic ratio over 90 % were kept, and segments below were discarded. The size of the frames going into the network was 25 ms and no overlap was used. To use vowels, and small segments where the audio signal may be considered stationary, was inspired by [21]. The same network structure as before, which is found in [5], was used, which can be seen in the appendix. In order to include the nasometry data the network had to be divided into two branches connecting at the end of the network as shown in figure 2.12.

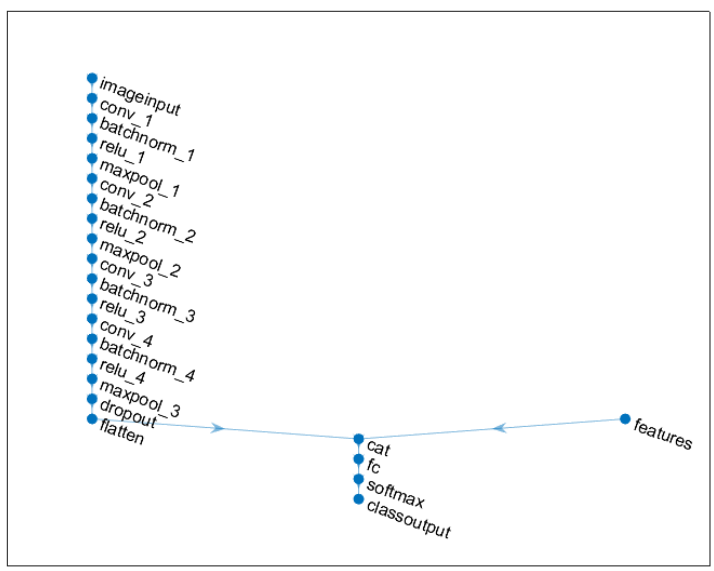

Figure 2.12: The network structure when including nasometry data.

As for the VPI data, the data sets were not evenly distributed between the classes. In order for this to not influence the training of the network, the number of files in the smallest class decided how many files were used for

training and validation. Roughly 80 % of the files in the smallest class were used for training and the same number of files for the other classes.

Chapter 3

Results

The results from the different tests are presented here. First how the methods in [5] worked on the new data and how better results could be obtained by changing a few settings. The results for classifying hypernasality are also presented.

3.1 Velopharyngeal Insufficiency

The Convolutional Neural Network implemented in [5] was used on the SC-data to obtain the results in figure 3.1 and 3.2. As can be seen both the frame-wise confusion plot and the file-wise do not show good results. The result is almost as good as guessing. A bad result was however expected since problems with the implementation had been identified before.

The VGGish implemented in [5] was used on the SC-data to obtain the results in figure 3.3 and 3.4. The results were not as good as the ones in [5], but better than the results from the CNN.

For the improved method the results for CNN and VGGish can be seen in figures 3.5, 3.6, 3.7 and 3.8. A clear improvement can be seen in the results with the adjustments made, for both CNN and VGGish. However, VGGish seems to give better results for VPI, than CNN.

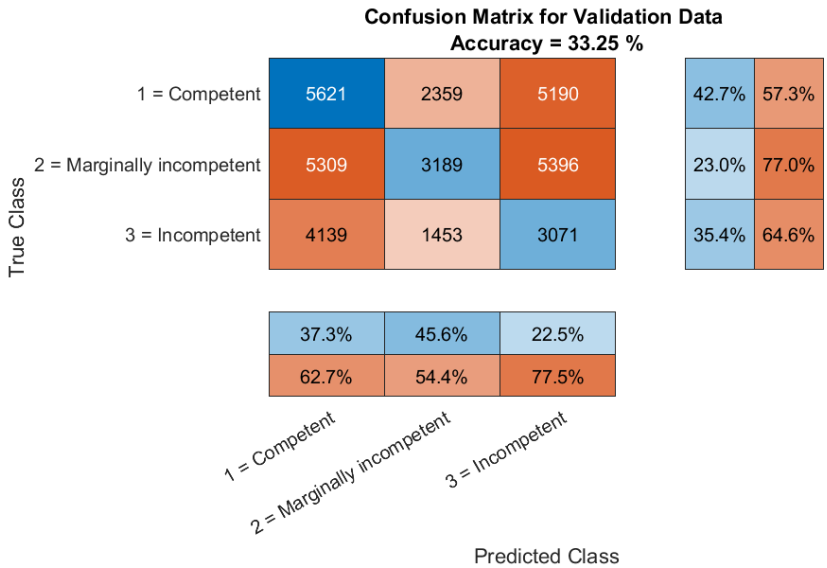

Figure 3.1: CNN: Frame-wise confusion plot for earlier method.

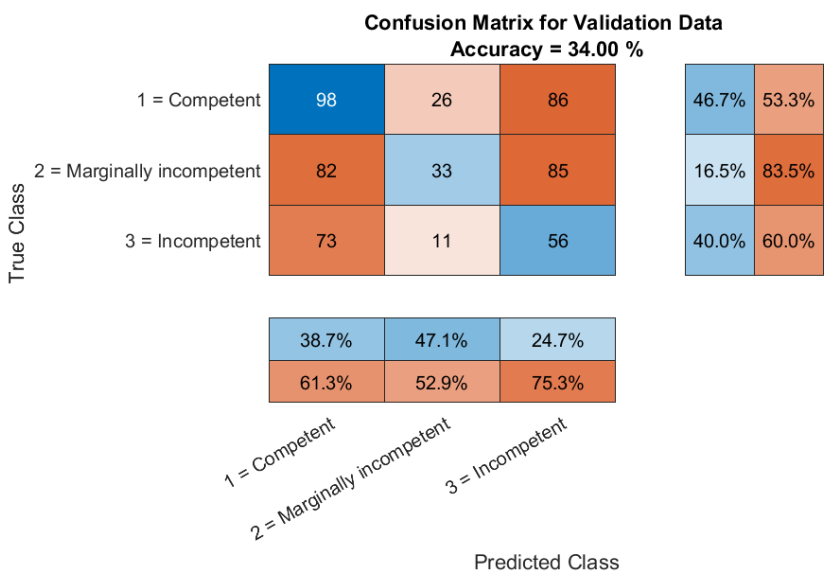

Figure 3.2: CNN: File-wise confusion plot for earlier method.

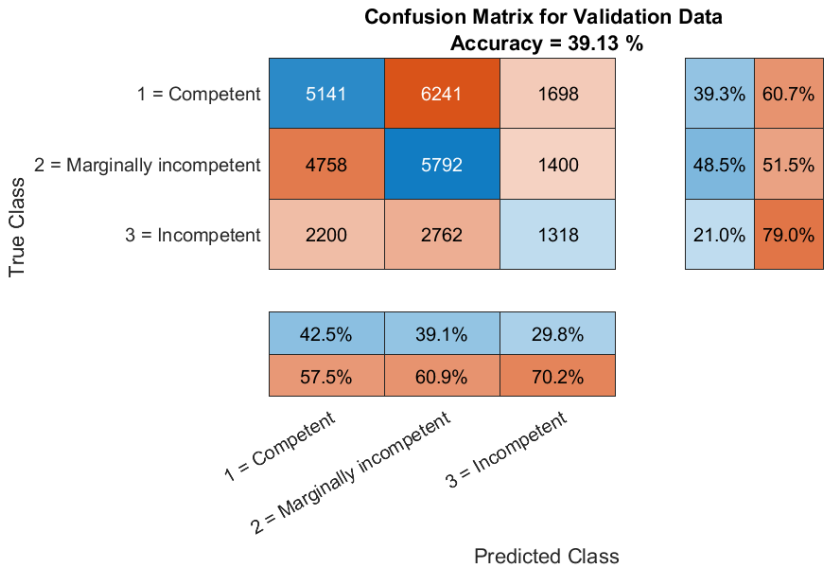

Figure 3.3: VGGish: Frame-wise confusion plot for earlier method.

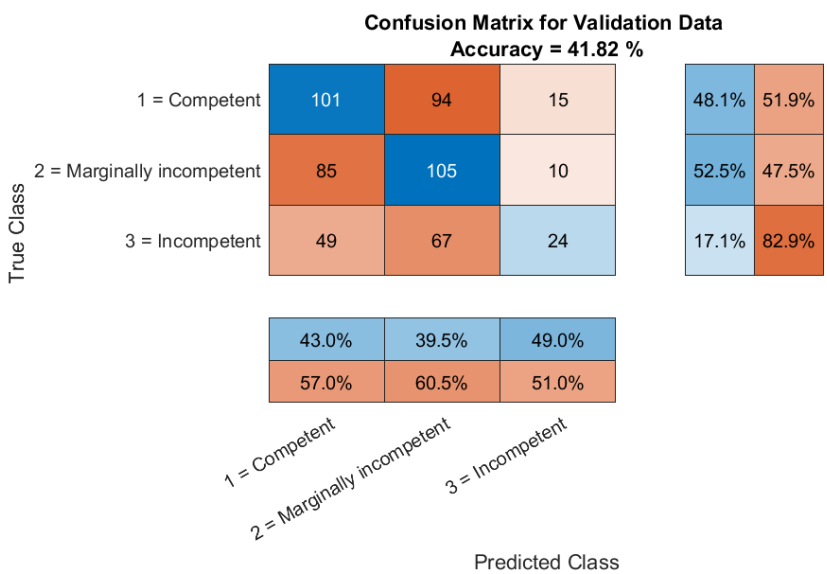

Figure 3.4: VGGish: File-wise confusion plot for earlier method.

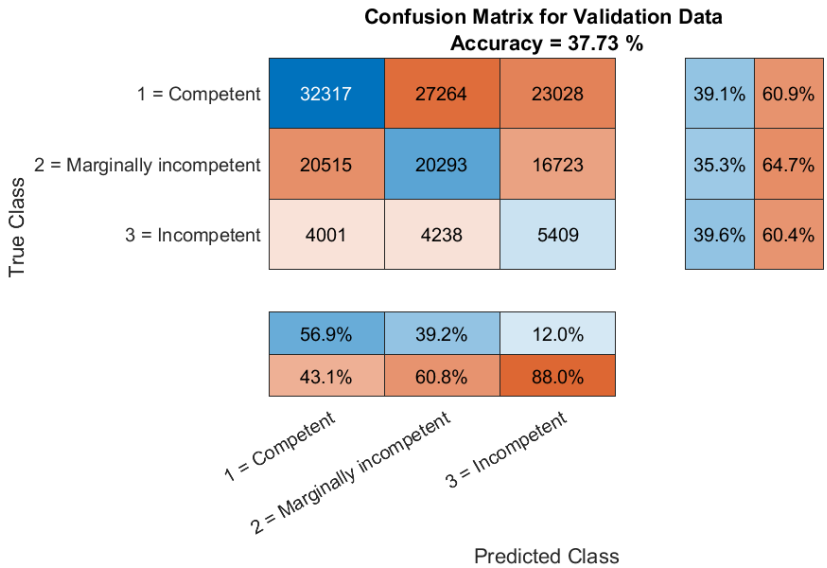

Figure 3.5: CNN: Frame-wise confusion plot for updated method.

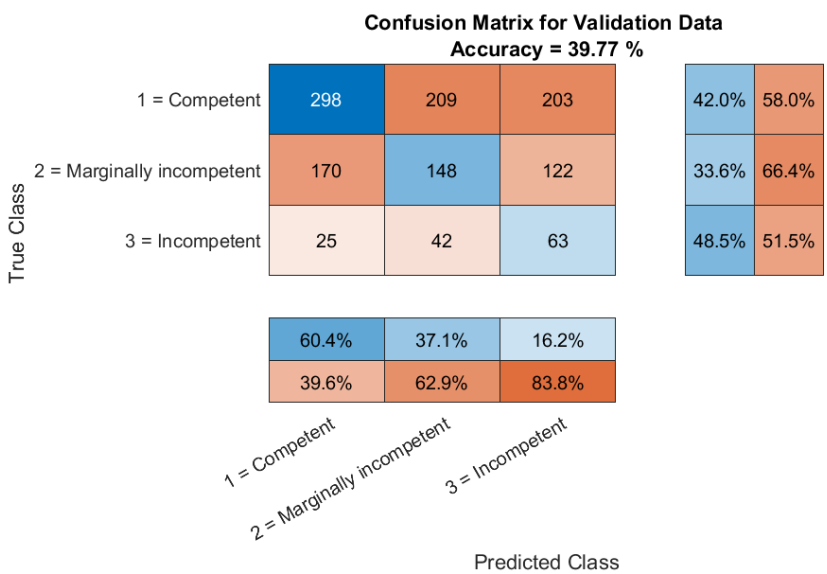

Figure 3.6: CNN: File-wise confusion plot for updated method.

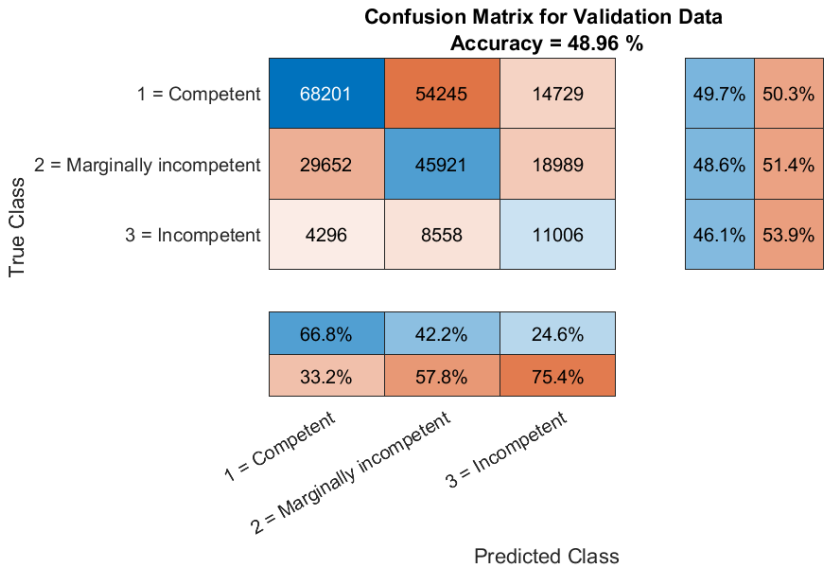

Figure 3.7: VGGish: Frame-wise confusion plot for updated method.

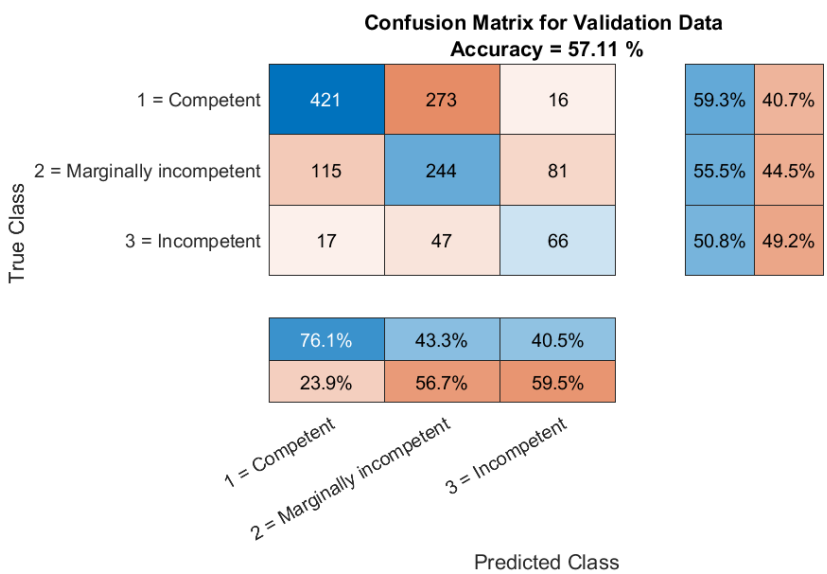

Figure 3.8: VGGish: File-wise confusion plot for updated method.

3.2 Hypernasality

Four different data sets were used for classifying hypernasality. The results when all files with the ratings 0-3 were used can be seen in figures 3.9 and 3.10. The results when nasometry data was used together with the audio files, with the ratings 0-3 can be seen in figures 3.11 and 3.12. It was also tested to use a binary data set. The result when a binary data set was used can be seen in figures 3.13 and 3.14. The binary data set was also tested with nasometry data and the result can be seen in figures 3.15 and 3.16. For all results 10 fold cross-validation has been used. Since the validation data is not evenly distributed, the most interesting results can be seen at the right side of the confusion plots as they are the percentages that has been correctly and incorrectly classified for each class.

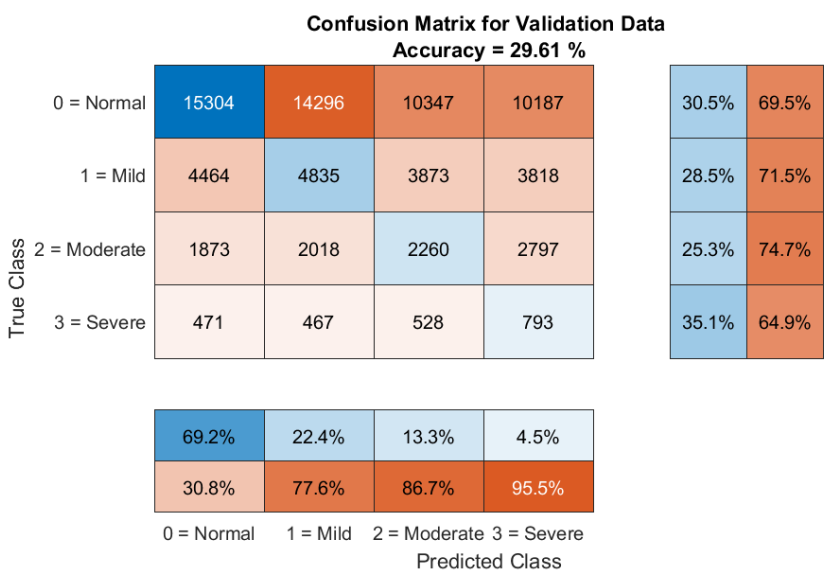

Figure 3.9: CNN: Frame-wise confusion plot for the four point scale data set without nasometry data.

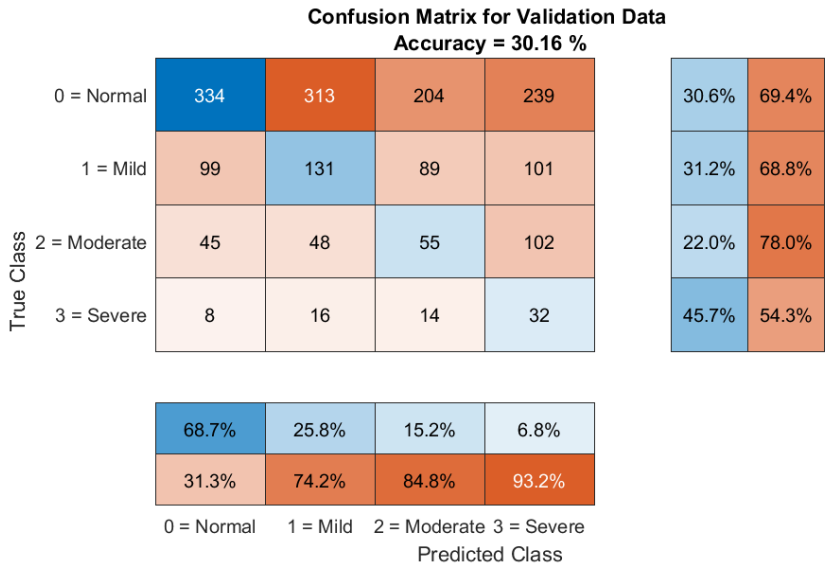

Figure 3.10: CNN: File-wise confusion plot for the four point scale data set without nasometry data.

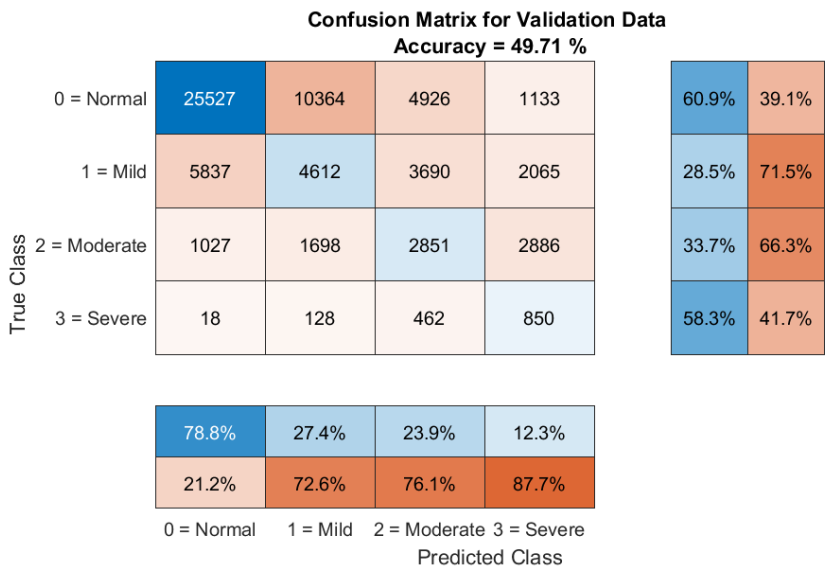

Figure 3.11: CNN: Frame-wise confusion plot for the four point scale data set with nasometry data.

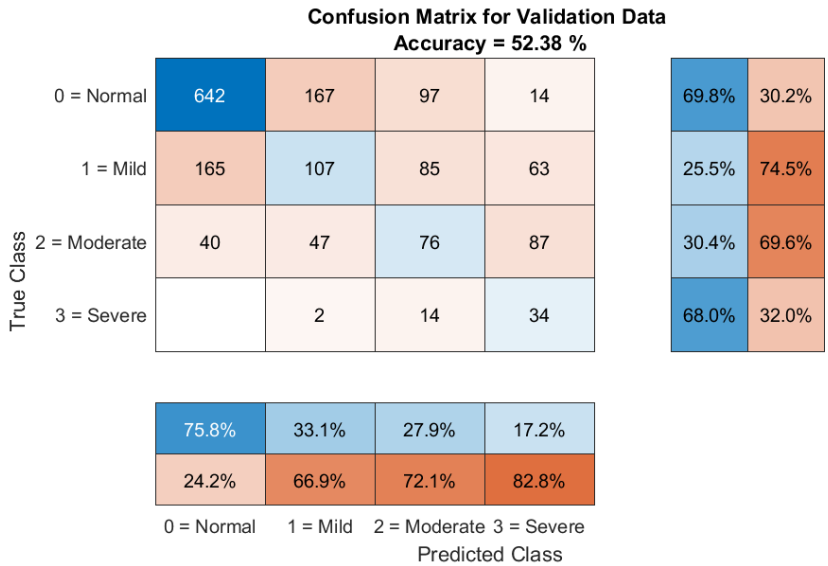

Figure 3.12: CNN: File-wise confusion plot for the four point scale data set with nasometry data.

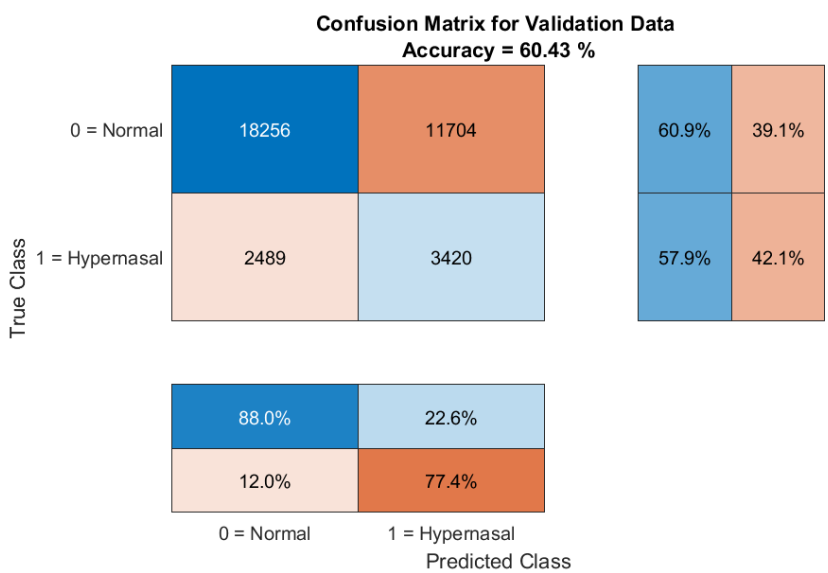

Figure 3.13: CNN: Frame-wise confusion plot for the binary data set without nasometry data.

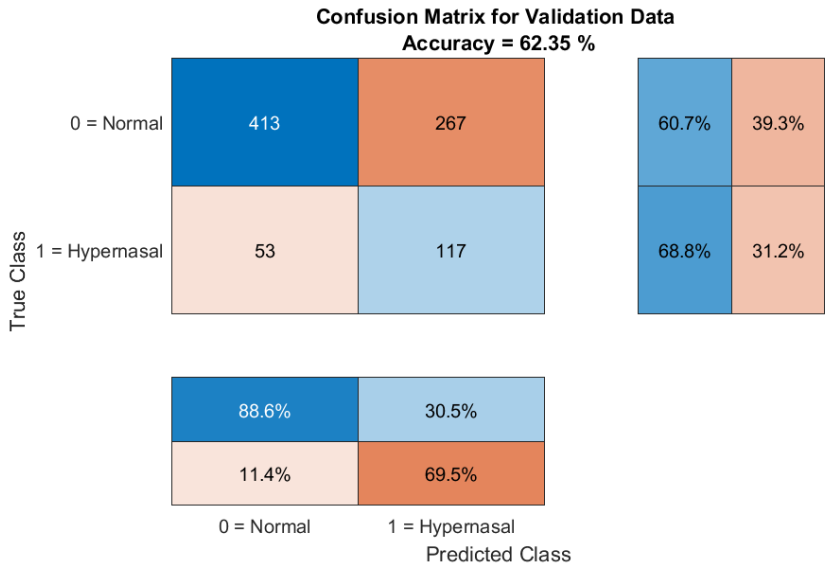

Figure 3.14: CNN: File-wise confusion plot for the binary data set without nasometry data.

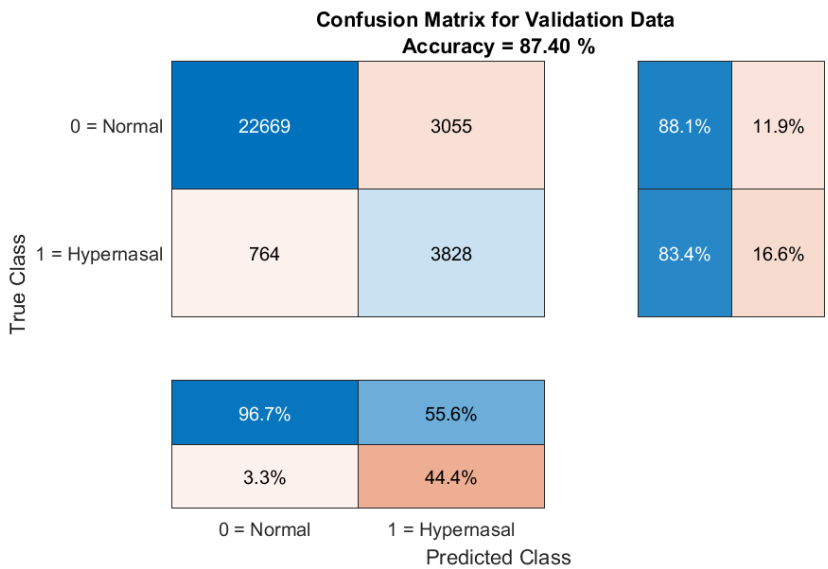

Figure 3.15: CNN: Frame-wise confusion plot for the binary data set with nasometry data.

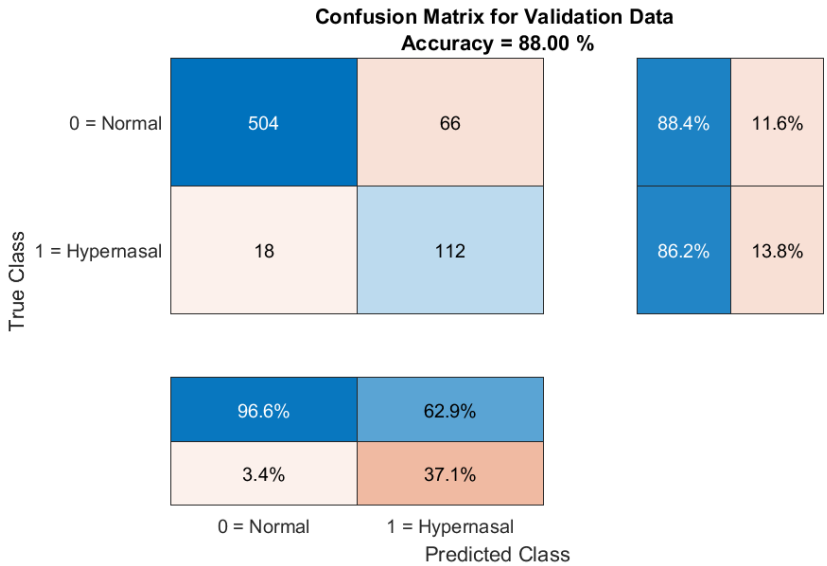

Figure 3.16: CNN: File-wise confusion plot for the binary data set with nasometry data.

Chapter 4

Discussion and Conclusion

In this part of the thesis the results are discussed and if the aim of the project was fulfilled. Ideas and recommendations for further work is also presented.

4.1 Discussion of Results

Overall the results achieved in this master's thesis were not as good as expected. A substantial part was probably due to the data used. Even if the languages in the SC-study are quite similar and the words being equivalent in relation to hypernasality, it may be hard to mix languages in this way. The evaluations from the speech pathologists also differ, especially for hypernasality. Even so the results from using deep learning does not perform better than the speech pathologists. However, since the ground truth is based on the ratings from the speech pathologists it may not even be possible, or at least very difficult, to perform better.

4.1.1 Velopharyngeal Insufficiency

The result when using the methods presented in [5] were really bad and the network were not able to classify VPI in a good way. A reason for this is probably that the pyAnnote program did not work as well for the data used in this thesis and were not able to separate the childrens speech in a good enough way. When the segmentation was done manually it could be concluded that many files were containing a lot of noise and other sounds, which could have influenced the segmentation. A better result was presented after

the segmentation had been done by hand, and a few things changed in the implementation. Even so the result was not satisfactory and the best accuracy found was 57.11 %, when using a VGGish network. The performance in [5] was 57.67% for the VGGish network, which is very similar. It seems that even if the children were both 5 and 10 years old in [5], the data was not as old as the one used in this master's thesis, and it only contained Swedish, the same performance was reached. The performance of the CNN can not be compared since it was found erroneous in [5], but in this work the VGGish network outperformed the CNN.

4.1.2 Hypernasality

The results when classifying hypernasality on a four point scale was not good. Without using nasometry data, only a file-wise accuracy of 30.16 %, was reached. However, with nasometry data, the file-wise accuracy was instead 52.38 %. The network especially had trouble with category 1=Mild and 2=Moderate, but was better at classifying 0=Normal and 3=Severe. The same can be seen for the speech pathologists, whom also had problems with the middle categories as seen in the methodology chapter. When the binary data set was used instead, a better result could be achieved. When using the binary data set together with the nasometry data, a file-wise accuracy of 88.00 %, was achieved, which is not bad. If the F1-score is calculated as in equation 1.2 from the probabilities found in the file-wise confusion plot in figure 3.16, the result is a F1-score of 0.874. This result is not that far away from 0.9485 achieved in [21]. However, the two studies differ a bit which makes it hard to compare them. The data set in [21] contains older children than in this study, and they are speaking Chinese, which differ significantly from the Nordic languages.

As can be seen in figure 2.11, the nasometry data from the IC-study is almost able to separate the data by itself. This is not the case for the SC-study, as seen in figure 2.10, but the same trend may be seen. It can be concluded that the nasometry data helps the network in order to classify hypernasality, but it could probably be used by itself as well to give good results.

4.2 Further Work

In order to improve the method of detecting VPI and hypernasality in cleft palate speech, one thing would be to improve the diarization. To manually segment out the children's speech is very time consuming, and a better way

would probably be needed. PyAnnote can be trained to perform better on specific tasks and could probably be used with better results. However, other methods could be tried as well and trained for the specific task of separating a child's voice from an adult. How the speech is recorded also plays a big role in this. In the data used in this project more than two people are sometimes present in the recording, there are also background noise and people speaking at the same time. If the data were more structured and without other noises and voices than the child and the speech pathologist it would be a simpler task. An idea would be that the speech pathologist record themselves beforehand to have a speech sample that an algorithm would recognize as the adult and should be segmented away, and then everything else would be the child speaking.

The 5 year old's speak very differently in the recordings. Some are laughing, some are speaking very quietly, some are not speaking and some are shouting, which makes it very hard to use the data and for the network to detect VPI and hypernasality. It would probably be a better idea to start by only using recordings from 10 year old's in order to get a good network structure that are able to detect VPI or hypernasality, that could later be used on 5 year old's. Data from 10 year old's are usually much easier to work with since the children can read by themselves and hence do not need to repeat sentences, or have their parents present in the recordings. They also speak more clearly overall and it is easier to detect differences in the speech since other speech deviations might not be present.

It is probably advisable to only focus on one language. Even if the languages used in this thesis are quite similar in many aspects, it adds an extra dimension since the perception of nasality may differ between languages.

Since vowels play a big role in classifying hypernasality a better way of extracting the vowels could be explored. Since some vowels, as for example [i], are more sensitive towards hypernasality it could be interesting to see if only extracting sensitive vowels would improve the network. It would also be interesting to see if only letting the children say different vowels would give better results. In [17], adults imitate cleft palate speech by only producing vowels, which when used in a neural network gives good results. It would probably be possible to test this on 10 year old's to see if the same good results can be achieved.

4.3 Conclusion

The conclusion that can be drawn from this master's thesis is that it might be possible to use deep learning in order to classify VPI and hypernasality, but in order to do so the recordings need to be of good quality and well segmented. To use older children than 5 year old's might be a way forward, in order to train networks that are able to found features related to VPI and hypernasality. To classify hypernasality, the results from this study indicate that it is a good idea to use nasometry data in order to help the classification.

Bibliography

[1] Audacity® software is copyright © 1999-2021 Audacity Team. It is free software distributed under the terms of the GNU General Public License. The name Audacity® is a registered trademark. URL `https://audacityteam.org/`.

[2] 1177-vårdguiden. Läppspalt, käkspalt, gomspalt - LKG-spalt. `https://www.1177.se/Skane/sjukdomar--besvar/mun-och-tander/mun-lappar-och-tunga/lappspalt-kakspalt-gomspalt---lkg-spalt/`, 2021. Accessed: 2022-02-14.

[3] A. Lohmander, E. Willadsen, C. Persson, G. Henningsson, M. Bowden, B. Hutters. Methodology for speech assessment in the Scandcleft project–an international randomized clinical trial on palatal surgery: experiences from a pilot study. *Cleft Palate–Craniofacial Journal*, 46(4), July 2009.

[4] Akademiska sjukhuset. Information om spaltmissbildningar. `https://www.akademiska.se/for-patient-och-besokare/ditt-besok/undersokning/lapp--kak--och-gomspalt/`. Accessed: 2022-02-14.

[5] J. Bluhme and T. Mamedov. *Exploring Deep Learning Approaches to Cleft Lip and Palate Speech*. Master's thesis, Lund University, Faculty of Engineering, 2021.

[6] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[7] K. Brunnegård and J. van Doorn. Normative data on nasalance scores for Swedish as measured on the Nasometer: Influence of dialect, gender,

and age. *Clinical Linguistics Phonetics*, 23(1):58–69, 2009. doi: 10.1080/02699200802491074.

[8] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[9] S. Howard and A. Lohmander. *Cleft Palate Speech: Assessment and Intervention*, chapter 9 and 11. Wiley-Blackwell, John Wiley Sons, Ltd., Publication, John Wiley Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK, 2011.

[10] K. Klintö, K. Brunnegård, C. Havstam, M. Appelqvist, E. Hagberg, A.-S. Taleman, and A. Lohmander. Speech in 5-year-olds born with unilateral cleft lip and palate: a Prospective Swedish Intercenter Study. *Journal of Plastic Surgery and Hand Surgery*, 53(5):309–315, 2019.

[11] J. Kreiman and D. Sidtis. *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Wiley-Blackwell, John Wiley Sons, Ltd., Publication, John Wiley Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK, 2011.

[12] V. C. Mathad, N. Scherer, K. Chapman, J. M. Liss, and V. Berisha. A deep learning algorithm for objective assessment of hypernasality in children with cleft palate. *IEEE Transactions on Biomedical Engineering*, 68(10):2986–2996, 2021.

[13] Mathworks, Inc. Specify Layers of Convolutional Neural Network. `https://se.mathworks.com/help/deeplearning/ug/layers-of-a-convolutional-neural-network.html`, . Accessed: 2022-05-10.

[14] Mathworks, Inc. Deep learning. `https://se.mathworks.com/discovery/deep-learning.html`, . Accessed: 2022-05-07.

[15] Mathworks, Inc. Harmonic Ratio. `https://se.mathworks.com/help/audio/ref/harmonicratio.html`, . Accessed: 2022-05-10.

[16] Mathworks, Inc. vggish. `https://se.mathworks.com/help/audio/ref/vggish.html`, . Accessed: 2022-05-07.

[17] N. Yogendran, A. Lohmander, F. Mehendale, D. Sell, and A. Tsanas. *Machine Learning-based Classification of Cleft Speech: CLeft-Associated Resonance Imitation Study (CLARIS)*. Edinburgh Medical SchoolUniversity of Edinburgh, 2021.

[18] L. Roberts. Understanding the Mel Spectrogram. `https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53`. Accessed: 2022-04-28.

[19] V. Skoog, O. Engström, M. Mani, M. Hakelius, and D. Nowinski. Läpp-käk-gomspalt (LKG). `https://www.internetmedicin.se/behandlingsoversikter/plastikkirurgi/lapp-kak-gomspalt-lkg/`, 2021. Accessed: 2022-06-09.

[20] R. E. Thomson and W. J. Emery. Chapter 5 - Time Series Analysis Methods. In R. E. Thomson and W. J. Emery, editors, *Data Analysis Methods in Physical Oceanography (Third Edition)*, pages 425–591. Elsevier, Boston, third edition edition, 2014. ISBN 978-0-12-387782-6.

[21] X. Wang, M. Tang, S. Yang, H. Yin, H. Huang, and L. He. Automatic Hypernasality Detection in Cleft Palate Speech Using CNN. *Circuits Syst Signal Process*, 38:3521–3547, 2019.

Appendix A

Table A1: CNN network architecture used. Almost the same as in [5]. Image input size VPI earlier: 96x64, VPI updated: 100x64 and hypernasality: 50x64.

Layer	**Type**	**Description**
1	Input	Image input layer
2	Conv	Convolutional layer, 12x3x3 filters, stride 2
3	Batch	Batch normalization layer
4	ReLU	ReLU layer
5	Pool	Maxpoolimg layer, 3x3 filter, stride 2
6	Conv	Convolutional layer, 24x3x3 filters, stride 2
7	Batch	Batch normalization layer
8	ReLU	ReLU layer
9	Pool	Maxpooling layer, 3x3 filter, stride 2
10	Conv	Convolutional layer, 48x3x3 filters, stride 1
11	Batch	Batch normalization layer
12	ReLU	ReLU layer
13	Conv	Convolutional layer, 48x3x3 filters, stride 1
14	Batch	Batch normalization layer
15	ReLU	ReLU layer
16	Pool	Maxpooling layer, 3x3 filter, stride 1
17	Drop	Dropout layer with 0.2 probability
18	FC	Fully connected layer with 3 neuron output
19	Output	Softmax classification layer

Table A2: VGGish network architecture.

Layer	**Type**	**Description**
1	VGGish	VGGish architecture, without the final layer
2	FC	Fully connected layer with 3 neuron output
3	Output	Softmax classification layer

Table A3: CNN training settings for the updated VPI method.

Parameter	**Setting**	**Description**
Mini batch size	512	Samples seen between weight updates
Epochs	20	Number of times the network sees all data
Optimizer	Adam	Minimization algorithm used
Loss function	Cross-entropy	Measure to quantify errors
Initial learning rate	0.001	Initial step size for optimizer
Learn rate drop factor	0.1	The factor which the learning rate drops during training
Learn rate drop period	10	The number of epochs between learning rate adjustment

Table A4: VGGish training settings for the updated VPI method.

Parameter	**Setting**	**Description**
Mini batch size	512	Samples seen between weight updates
Epochs	3	Number of times the network sees all data
Optimizer	Adam	Minimization algorithm used
Loss function	Cross-entropy	Measure to quantify errors
Initial learning rate	0.001	Initial step size for optimizer
Learn rate drop factor	0.1	The factor which the learning rate drops during training
Learn rate drop period	2	The number of epochs between learning rate adjustment

Table A5: CNN training settings for hypernasality.

Parameter	**Setting**	**Description**
Mini batch size	64	Samples seen between weight updates
Epochs	20	Number of times the network sees all data
Optimizer	Adam	Minimization algorithm used
Loss function	Cross-entropy	Measure to quantify errors
Initial learning rate	0.001	Initial step size for optimizer
Learn rate drop factor	0.1	The factor which the learning rate drops during training
Learn rate drop period	10	The number of epochs between learning rate adjustment