diagnostics

MDPI

Article

Automatic Detection System for Velopharyngeal Insufficiency Based on Acoustic Signals from Nasal and Oral Channels

Yu Zhang [1], Jing Zhang [1], Wen Li [1], Heng Yin [2] and Ling He [1,*]

[1] College of Biomedical Engineering, Sichuan University, Chengdu 610065, China; zhangyu5@stu.scu.edu.cn (Y.Z.); jing_zhang@scu.edu.cn (J.Z.); liwen1@stu.scu.edu.cn (W.L.)
[2] West China Hospital of Stomatology, Sichuan University, Chengdu 610041, China; phoebeyin@126.com
* Correspondence: ling.he@scu.edu.cn

Abstract: Velopharyngeal insufficiency (VPI) is a type of pharyngeal function dysfunction that causes speech impairment and swallowing disorder. Speech therapists play a key role on the diagnosis and treatment of speech disorders. However, there is a worldwide shortage of experienced speech therapists. Artificial intelligence-based computer-aided diagnosing technology could be a solution for this. This paper proposes an automatic system for VPI detection at the subject level. It is a non-invasive and convenient approach for VPI diagnosis. Based on the principle of impaired articulation of VPI patients, nasal- and oral-channel acoustic signals are collected as raw data. The system integrates the symptom discriminant results at the phoneme level. For consonants, relative prominent frequency description and relative frequency distribution features are proposed to discriminate nasal air emission caused by VPI. For hypernasality-sensitive vowels, a cross-attention residual Siamese network (CARS-Net) is proposed to perform automatic VPI/non-VPI classification at the phoneme level. CARS-Net embeds a cross-attention module between the two branches to improve the VPI/non-VPI classification model for vowels. We validate the proposed system on a self-built dataset, and the accuracy reaches 98.52%. This provides possibilities for implementing automatic VPI diagnosis.

Keywords: velopharyngeal insufficiency; speech disorder; automatic diagnosis; deep learning

check for updates

Citation: Zhang, Y.; Zhang, J.; Li, W.; Yin, H.; He, L. Automatic Detection System for Velopharyngeal Insufficiency Based on Acoustic Signals from Nasal and Oral Channels. *Diagnostics* **2023**, *13*, 2714. https://doi.org/10.3390/diagnostics13162714

Academic Editor: Jae-Ho Han

Received: 8 July 2023
Revised: 18 August 2023
Accepted: 19 August 2023
Published: 21 August 2023

1. Introduction

Velopharyngeal insufficiency and/or incompetency (VPI) refers to abnormal palatopharyngeal function. The airflow passage between the soft palate and the pharyngeal walls (or adenoids in children) does not close completely [1]. This defect severely affects the patient's daily life. VPI causes characteristic speech disorders, including hypernasality and nasal air emission [2]. They reduce the clarity of the patient's speech. It causes situational difficulty and emotional impacts, which affects normal interpersonal communication [3]. Additionally, VPI is closely related to swallowing disorders, which affects daily eating [4]. These symptoms seriously affect the physical and mental health of patients.

According to the causes of the disease, VPI can be classified as congenital VPI or acquired VPI [5]. Congenital developmental malformations, such as cleft palate [6] and congenital myotonic dystrophy [7], can lead to congenital VPI. In terms of acquired VPI, tonsillectomy [8], adenoidectomy [9], palatal sail shortening, or trauma [10] can cause it. VPI treatment requires continuous therapy. Approximately 5–35% of cleft palate patients still suffer from VPI after palatopharyngeal repair surgery [11]. This percentage even reaches 40% if it is determined strictly [12]. Treatment of VPI requires surgery or prolonged voice training. The timely diagnosis of VPI is important for the early treatment of patients and the acquisition of normal speech and language skills for children less than 3 years old [13].

Clinical assessment of palatopharyngeal function can be performed using instruments that allow direct visualization of the palatopharyngeal closure status. These include nasoendoscopy [14], multiview videofluoroscopy [15], and medical imaging methods [16–19]. Researchers have tried to find some velopharyngeal anatomy or velar shape parameters obtained by these instruments to predict whether the prognostic patients had VPI speech disorders or not [20,21]. The majority of patients with VPI are children. These methods are not child-friendly due to their intrusiveness, the involvement of radiation, and the requirement for patients to be highly cooperative.

The noninvasive approach generally relies on a speech-language therapist's (SLT) assessment of certain phenomena or intermediate data to diagnose hypernasality or nasal air emission caused by VPI. The mirror-fogging test is used to detect nasal air emission [22]. However, it can be affected by resistance of the nasal airways and only assesses nasal breathing without speech production [23]. Nasalance scores determined by Nasometer are commonly used in clinical practice to supplement assessment of hypernasality [24]. The scores represent the energy ratio of the acoustic signals of the nasal and oral channels. Researchers have noted that nasalance score does not have a fixed evaluation criterion for different languages [23]. The use of the Nasometer for determining nasalance scores exclusively relies on the energy of the acoustic signals, inevitably leading to the omission of specific speech perception information [25].

Clinical assessment using speech perception mostly relies on well-trained SLTs for subjective diagnosis of patient speech [26]. The number of SLTs is low, and training experienced SLTs requires a certain amount of time and money. End-to-end VPI detection algorithms, which employ speech signal processing and deep learning techniques, can effectively provide detection results that assist in a clinical VPI diagnosis. These algorithms eliminate the need for additional complex analysis by SLTs. They are both economical and convenient, effectively addressing the issue of late VPI detection in secondary care facilities or underdeveloped medical areas where speech-language therapists may not be readily available.

The current research on speech-based computer-aided VPI detection algorithms is focused on the automatic detection of speech disorders caused by VPI. In terms of hypernasality speech, the extra nasal resonance is present [27]. The nasal formant has been shown to be present around F1 [28]. The extraction algorithms of nasal formant [29,30] have been studied, and formants related characteristic parameters are utilized in automatic hypernasality detection. They contain group delay function-based acoustic measure (GDAM, the ratio of the absolute value at F1 to that at F2 in group delay spectrum) [31], the cross-correlation value of original speech signals and modified speech signals after pole-defocusing [32], the vowel spectral area (VSA) [33], and spectrum-based features [34,35]. The above methods with nasal formant or formants-related parameters are susceptible to age, gender, and noise. Recently, deep learning methods have been used in the study of automatic hypernasality classification, such as deep RNN [36], CNN [37], and improved BLSTM [38]. To solve the problem of sparse hypernasality speech data, researchers [39,40] attempted to use automatic speech recognition models trained by normal speech for the diagnosis of hypernasality in children. However, the validation datasets do not include data from adult VPI patients. In terms of nasal air emission, there are few automatic detection methods. Nasal air emission and hypernasality due to VPI can occur separately or together [41]. There is also a lack of methods that directly detect VPI automatically.

This paper proposes an automatic VPI detection system that operates at the subject level. The data are the acoustic signals collected from the nasal and oral channels. VPI/non-VPI automatic classification methods at the phoneme level are proposed for consonants and vowels. The results at the phoneme level for the subjects are fused to obtain the VPI/non-VPI detection result at the subject level. The above methods can assist clinicians in VPI diagnosis. Our main contributions are summarized as follows:

(1) An automatic VPI/non-VPI detection system at the subject level based on speech is proposed in this paper. The system takes into account distinctive articulatory

symptoms caused by VPI, nasal air emission, and hypernasality. For this purpose, the detection of VPI/non-VPI at the subject level is proposed by integrating the symptomatic manifestations at the phoneme level.

(2) VPI causes a change in the propagation path of the airflow through the vocal tract, which affects the acoustic signals radiated from the nasal and oral cavities. For unvoiced consonants, the power spectral density ratio (PSDR) is calculated to indicate airflow leakage to the nasal cavity relative to the oral cavity. This paper proposes relative prominent frequency description and relative frequency distribution features based on PSDR. They are extracted to characterize the perceived acoustic signals radiated from the nasal cavity relative to those from the oral cavity.

(3) Mathematical models for VPI patients and non-VPI controls on vowel articulation are established in this paper. Based on the discrepancy between VPI patients and non-VPI controls shown in the models, a cross-attention residual Siamese network (CARS-Net) is proposed for VPI/non-VPI classification at the phoneme level for vowels. A cross-attention module is proposed that is embedded in CARS-Net to enhance the ability to extract the discriminating features for VPI and non-VPI classification at the phoneme level for vowels.

2. Materials and Methods

2.1. Materials

2.1.1. The Collected Phonemes

The specific clinical structural speech disorders caused by VPI are nasal air emission and hypernasality [2]. The vibration source of voiced phoneme production comes from the vocal cord, which is located in front of the palatopharynx in the path of airflow propagation. When patients with VPI produce nonnasalized vowels, some of the airflow carrying the vocal cord wave propagates to the nasal cavity and causes nasal resonance, resulting in hypernasality. In contrast, the production of unvoiced consonants relies on the friction between the airflow and the various articulatory parts of the vocal tract, most of which appear behind the palatopharynx [42]. In this case, the airflow leakage to the nasal cavity does not cause strong nasal resonance. However, compared to voiced phonemes, the unvoiced consonants included in unvoiced phonemes have continuous airflow overflow and reduce the effect of nasal resonance, which results in the nasal air emission symptoms [43].

Nasal air emission symptoms caused by VPI arise from specific types of consonants: plosives, affricates, and fricatives [44]. The airflow exhaled from the lungs during unvoiced consonant production is the source of the nasal air emission symptoms caused by VPI. The plosives and affricates are divided into aspirated and nonaspirated phonemes, depending on the relative size of the airflow delivery. In contrast to the nonaspirated phonemes, the aspirated phonemes produce a distinct airflow during articulation. The fricatives are also pronounced with a distinct airflow output. In this work, the aspirated unvoiced consonants are collected into the dataset (/p/, /t/, /k/, /q/, /c/, /h/, /x/, /sh/, /f/).

Hypernasality symptoms caused by VPI are mainly detected in the articulation of vowels [45]. When nasalized vowels are pronounced normally, the velopharynx is open [46]. To highlight the difference between VPI and non-VPI speech on vowels, nonnasalized vowels are chosen, which have almost complete closure of the palatopharynx during normal articulation. In this work, four nonnasalized vowels are considered (/a/, /e/, /i/, /u/).

2.1.2. Dataset

The data used in this work were collected from volunteers recruited by the Cleft Lip Unit of West China Hospital of Stomatology, Sichuan University. A total of 89 patients with VPI and 46 controls without VPI were included in the dataset. The participants are from 10 provinces in China and aged from 4 to 45. And the average ages (mean ± SD) of the VPI and non VPI groups are 18.64 ± 7.58 and 13.04 ± 7.04, respectively. In terms of sex, there are 66 females and 69 males. Among them, the VPI group consists of 43 females and 46 males, and the non-VPI group consists of 23 females and 23 males. There are 4860 phonemes in the dataset.

The Nasometer II 6450 (kayPENTAX, State of New Jersey, USA) is the acquisition device for the dataset. The two microphones carried by this device are separated by a plate, allowing simultaneous recording of the acoustic signals radiating from the nasal and oral cavities. The nasal channel records the sound of a microphone placed close to the nasal cavity, and the oral channel corresponds to the sound recorded by the microphone placed in front of the oral cavity. The sampling rate is 11,025 Hz.

2.2. Overview of the Automatic VPI Detection Method

An automatic VPI/non-VPI detection system at the subject level is proposed in this work. As shown in Figure 1, the process is divided into two steps.

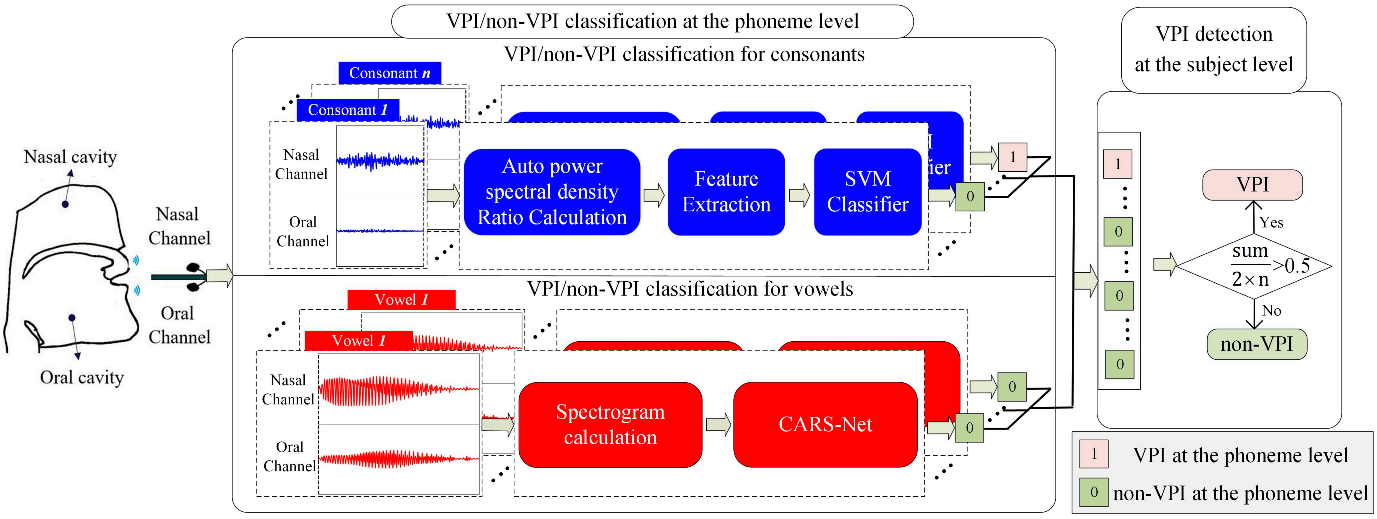
VPI/non-VPI classification at the phoneme level
VPI/non-VPI classification for consonants
Consonant n
Consonant 1
Nasal cavity
Nasal Channel
Oral Channel
Oral cavity
Nasal Channel
Oral Channel
Auto power spectral density Ratio Calculation
Feature Extraction
SVM Classifier
1
0
VPI/non-VPI classification for vowels
Vowel 1
Vowel 1
Nasal Channel
Oral Channel
Spectrogram calculation
CARS-Net
0
0
VPI detection at the subject level
1
0
0
0
VPI
Yes
sum / 2×n >0.5
No
non-VPI
1 VPI at the phoneme level
0 non-VPI at the phoneme level

Figure 1. Overview of VPI detection at the subject level.

Step 1: VPI/non-VPI classification at the phoneme level.

Due to the different effects of VPI on the pronunciation of unvoiced consonants and vowels, different classification methods are proposed for consonants and vowels, as described below.

(1) VPI/non-VPI classification at the phoneme level for consonants: The relative prominent frequency description and relative frequency distribution are extracted based on the power spectral density ratio sequence. They are combined with a support vector machine classifier to implement the VPI/non-VPI consonant classification model.

(2) VPI/non-VPI classification at the phoneme level for vowels: A cross-attention residual Siamese network (CARS-Net) is proposed in this paper to achieve end-to-end VPI/non-VPI speech classification at the phoneme level for vowels.

Step 2: Automatic VPI detection at the subject level by the voting method.

The VPI/non-VPI classification results for all phonemes of each subject are aggregated through voting to obtain the VPI detection results at the subject level. A subject is classified as a VPI patient if more than half of the participating phonemes are classified as VPI speech.

2.3. VPI/Non-VPI Classification Methods at the Phoneme Level for Consonants

The proposed automatic VPI/non-VPI speech classification method for consonants is shown in Figure 2. The power spectral density of the acoustic signals radiated from the nasal cavity and oral cavity is approximated by that of the acoustic signals collected by the equipment. Relative prominent frequency description (*RPFD*) features and relative frequency distribution (*RFD*) features between the acoustic signals of nasal and oral channels are obtained. They are extracted to model the relative sounds produced by airflow leakage to the nasal cavity. SVM classifier is used to discriminate unvoiced consonant production with and without symptoms caused by VPI.

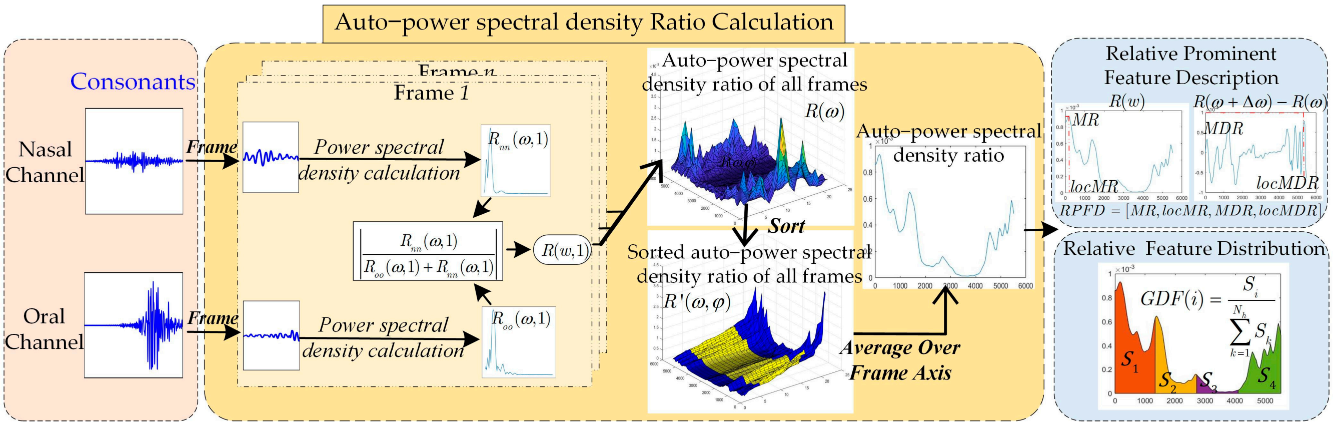
Auto−power spectral density Ratio Calculation
Consonants
Nasal Channel
Oral Channel
Frame
Frame
Frame n
Frame 1
Power spectral density calculation
$R_{nn}(\omega,1)$
$\frac{R_{nn}(\omega,1)}{R_{oo}(\omega,1)+R_{nn}(\omega,1)}$
$R(w,1)$
Power spectral density calculation
$R_{oo}(\omega,1)$
Auto−power spectral density ratio of all frames
$R(\omega)$
Sort
Sorted auto−power spectral density ratio of all frames
$R'(\omega,\varphi)$
Average Over Frame Axis
Auto−power spectral density ratio
Relative Prominent Feature Description
$R(w)$
$R(\omega+\Delta\omega)-R(\omega)$
MR
locMR
MDR
locMDR
$RPFD = [MR, locMR, MDR, locMDR]$
Relative Feature Distribution
$GDF(i) = \frac{S_i}{\sum_{k=1}^{N_b} S_k}$
S_1
S_2
S_3
S_4

Figure 2. Feature extraction process for consonants.

2.3.1. Power Spectral Density of Radiated Signals from the Nasal Cavity and Oral Cavity

The power spectral density ratio of the acoustic signals of the nasal channel and oral channel is calculated to discriminate unvoiced consonants with and without nasal air emission symptoms caused by VPI. The acoustic signals recorded by the microphones do not only contain the acoustic signal radiated from one cavity, even if there is a separator plate blocking the acoustic signal transmission. This subsection describes the mathematical derivation of the power spectral density of the collected acoustic signals of one channel to approximate the power spectral density of the radiated acoustic signals from the corresponding cavity.

The consonant segments of two-channel acoustic signals collected by microphones in front of the nasal and oral cavities are divided into frames with a frame length of 20 ms and a frame shift of 6 ms. The i-th frame signals of the oral channel and the nasal channel are denoted as $x_{o_i}(t)$ and $x_{n_i}(t)$.

Assuming the signal attenuation of the microphone with respect to the other channel is a linear attenuation; the relationship between the collected signals and the radiated signals is as shown in the following equations.

$$\begin{cases} x_{o_i}(t) = o_i(t) + \alpha \cdot n_i(t) \\ x_{n_i}(t) = \beta \cdot o_i(t) + n_i(t) \end{cases} \tag{1}$$

where $o_i(t)$ and $n_i(t)$ represent the signals radiated from the oral and nasal cavities, respectively. β and α are the attenuation coefficients of the baffle for oral and nasal channel acoustic signals, respectively, and β can be regarded as approximately the same as α.

The pronunciation of the unvoiced consonants is based on the friction between the airflow and the vocal organs. When patients with VPI produce unvoiced consonants, the friction factor generated by the airflow overflowing the nasal cavity is different from that of the oral cavity. Therefore, the frictional sounds radiated from each of the two cavities are regarded as uncorrelated signals in this paper. Then, the autocorrelation functions of $x_{o_i}(t)$ can be derived as follows.

$$\begin{aligned} r_{oo}(m,i) &= \sum x_{o_i}(t) x_{o_i}(t+m) \\ &= \sum \left\{ o_i(t)o_i(t+m) + \alpha o_i(t)n_i(t+m) + \alpha n_i(t)o_i(t+m) + \alpha^2 n_i(t)n_i(t+m) \right\} \\ &= \sum o_i(t)o_i(t+m) + \sum \alpha^2 n_i(t)n_i(t+m) \\ &= s_{oo}(m,i) + \alpha^2 s_{nn}(m,i) \end{aligned} \tag{2}$$

Similarly, the autocorrelation functions of $x_{n_i}(t)$ can be obtained as,

$$r_{nn}(m,i) = s_{nn}(m,i) + \alpha^2 s_{oo}(m,i) \tag{3}$$

where $r_{oo}(m,i)$ and $r_{nn}(m,i)$ represent the autocorrelation functions of $x_{o_i}(t)$ and $x_{n_i}(t)$, respectively. And $s_{oo}(m,i)$ and $s_{nn}(m,i)$ represent the autocorrelation functions of $o_i(t)$ and $n_i(t)$, respectively.

Since the value of α^2 tends toward zero more than α, we can assume α^2 equals zero. According to well-known Wiener–Khinchin theorem, the power spectral density of the radiated signals from the nasal and oral cavities can be obtained by applying the Fourier transform to Equations (2) and (3), as shown in the following equations.

$$\begin{cases} R_{nn}(\omega,i) = FFT\big(s_{nn}(m,i) + \alpha^2 s_{oo}(m,i)\big) \approx\ FFT(s_{nn}(m,i)) = S_{nn}(\omega,i) \\ R_{oo}(\omega,i) = FFT\big(s_{oo}(m,i) + \alpha^2 s_{nn}(m,i)\big) \approx\ FFT(s_{oo}(m,i)) = S_{oo}(\omega,i) \end{cases} \tag{4}$$

where R_{oo} and R_{nn} represent the power spectral densities of $x_{o_i}(t)$ and $x_{n_i}(t)$, respectively. And S_{oo} and S_{nn} represent the power spectral densities of $o_i(t)$ and $n_i(t)$, respectively.

Then, the modulus of the approximate power spectral density ratio of the radiated signals from the nasal and oral cavities of the i-th frame $R(\omega,i)$ can be obtained, as shown in the following equation.

$$R(\omega,i) = \left|\frac{S_{nn}(\omega,i)}{S_{oo}(\omega,i) + S_{nn}(\omega,i)}\right| = \left|\frac{R_{nn}(\omega,i)}{R_{oo}(\omega,i) + R_{nn}(\omega,i)}\right| \tag{5}$$

where $|\cdot|$ represents the modulo operation.

2.3.2. Calculation of the Average Power Spectral Density Ratio of the Acoustic Signals Radiated from Nasal and Oral Cavities

To remove outliers caused by random noise, the ratios $R(\omega,\varphi)$ of specifical frequencies ω ordered on the frame axis from smallest to largest and are denoted as $R'(\omega,\gamma)$, where γ is the ranking position of the specified frequency. The average of the values located between the upper quartiles and lower quartiles is taken as the average power spectral density ratio, as shown in (6).

$$R(\omega) = \frac{\sum_{k=[\frac{N_f}{4}]}^{[\frac{N_f*3}{4}]} R'(\omega,k)}{\left[\frac{N_f*3}{4}\right] - \left[\frac{N_f}{4}\right] + 1} \tag{6}$$

where N_f represents the number of frames and $[\cdot]$ represents rounding up operation.

2.3.3. VPI Consonant Production Feature Extraction

(1) Relative prominent frequency description between the acoustic signals of nasal and oral channels

The generation of aspirated consonants is directly caused by the friction between the airflow and the gap of the cavity. The intensity of the sounds is positively correlated with the gas dynamic pressure. When patients with VPI produce unvoiced consonants, the airflow overflows into the nasal cavity. Assuming a constant cavity gap and lung dynamics, the relative gas dynamic pressure in the nasal cavity is increased, while the oral gas dynamic pressure is relatively weakened.

In this paper, we extract the maximums of the spectral density ratio (MR) and its first-order differential (MDR) to characterize the relative air flow to produce acoustic signals, as shown in (7)–(9).

$$MR = \max(R(\omega)) \tag{7}$$

$$MDR = \max(R(\omega + \Delta\omega) - R(\omega)) \tag{8}$$

$$\Delta\omega = \frac{f_s}{N_s} \tag{9}$$

where f_s stands for the sampling rate in the time domain, and N_s is the number of frequency sampling points. In this paper, N_s is set to 256.

In addition, the sound quality is related to the shapes of gaps and cavities, which are different in the nasal and oral channels. Therefore, the two cavities produce different frictional noises. The locations of MR and MDR are extracted to reflect the differences in the frequency domain.

$$locMR = arg\max_{\omega}(R(\omega)) \tag{10}$$

$$locMDR = arg\max_{\omega}(R(\omega + \Delta\omega) - R(\omega)) \tag{11}$$

The above three features are concatenated to form a relative prominent frequency description ($RPFD$), as shown in the following formula.

$$RPFD = [MR, locMR, MDR, locMDR] \tag{12}$$

(2) Relative frequency distribution between the acoustic signals of nasal and oral channels

Aspirated consonants are produced by turbulence, which is generated by friction between the airflow and the cavity gap. The different gap shapes of the nasal and oral cavities create different tuning effects. As a result, the frequency band distributions of the signals radiated from the two cavities are different. When patients with VPI produce aspirated consonants, compared with patients without VPI, the spectral density ratio is enhanced compared to the value at the frequency of the signal radiated from the nasal cavity. This is reflected by the difference in the distributions of the power spectral density ratio.

The frequency band of the power spectral density is linearly divided into N_h subbands. The percentage of each subband area to the total area of the frequency band of the power spectral density ratio is calculated to reflect the relative frequency distribution (RFD), as shown in (13).

$$RFD(i) = \frac{\sum_{\omega=\frac{(i-1)*fs}{N_h*2}}^{\frac{i*fs}{N_h*2}} R(\omega)}{\sum_{\omega=0}^{fs/2} R(\omega)} \tag{13}$$

where N_h is the total number of frequency sub-bands. In this work, N_h is set to 4.

2.4. VPI/Non-VPI Classification Methods at the Phoneme Level for Vowels

2.4.1. CARS-Net Proposed for VPI/Non-VPI Vowels Classification

In this paper, a cross-attention residual Siamese network (CARS-Net) is proposed to achieve automatic VPI/non-VPI classification for vowels. Based on the articulation principle of VPI speech, two-channel acoustic signals from oral and nasal cavities are collected for automatic classification. Compared with a single-input network, the Siamese network structure has dual inputs and is suitable for extracting differences in acoustic signals from the oral and nasal cavities. The network structure proposed in this paper is shown in Figure 3.

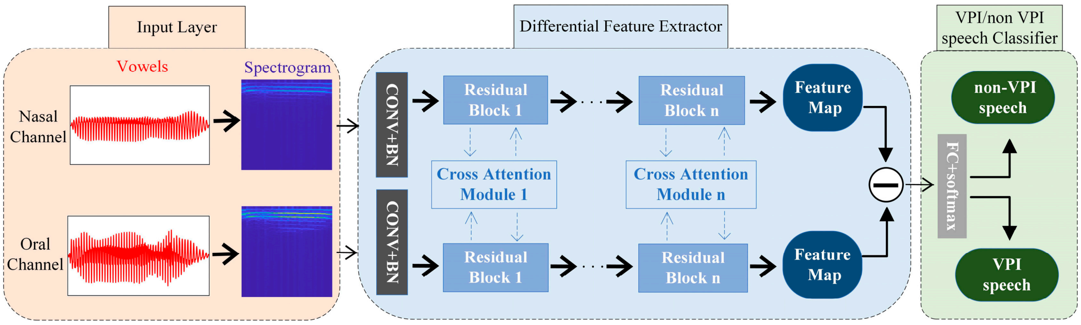

Figure 3. Cross-attention residual Siamese network for VPI/non-VPI classification for vowels.

As shown, CARS-Net contains three parts: the input layer, difference feature extractor (DFE), and VPI/non-VPI speech classifier. First, the acoustic signals of vowels are collected by two microphones in front of the oral cavity and nasal cavity. They are transformed into spectrograms as the network inputs. Then, the inputs go through the DFE to produce the difference feature map. The DFE contains two branching networks for the two inputs. The two branching networks are linked by the cross-attention module proposed in this paper. Finally, the difference feature map is fed into the VPI/non-VPI speech classifier for automatic classification. The next four sections describe the details of the three parts and loss function for training the network.

2.4.2. Input Layer

The oral and nasal channels of vowels are transformed into spectrograms by framing and short-term Fourier transform operations, and they are then used as inputs to the network.

A schematic diagram of the vowel production process is shown in Figure 4. $e(t)$ represents the vocal cord wave signal generated at the vocal cords. The airflow carrying the vocal cord waves diverges at the pharyngeal wall, partly to the oral cavity and partly to the nasal cavity, denoted as $e_{n_i}(t)$ and $e_{o_i}(t)$, respectively. $h_{n_i}(t)$ and $h_{o_i}(t)$ represent the system response generated by nasal and oral cavity, respectively.

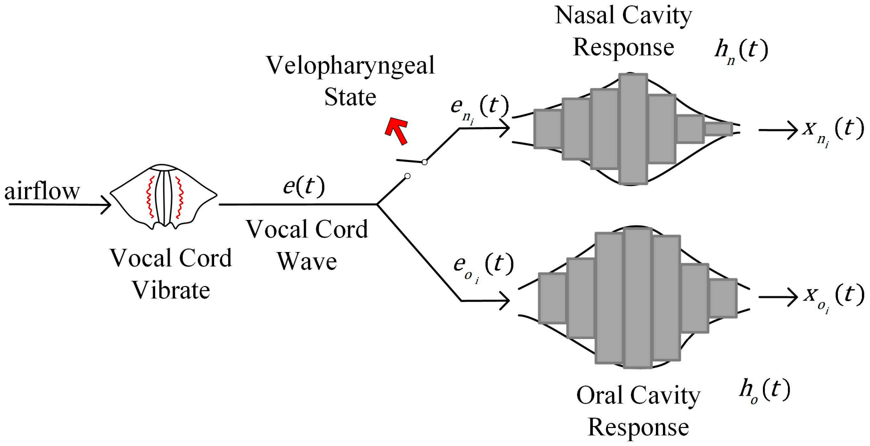

Figure 4. Schematic diagram of the vowel pronunciation with VPI or not.

There is a discrepancy between the two channel signals for VPI patients and controls. When a person without velopharyngeal insufficiency (VPI) vocalizes, the switch shown in the figure is open, indicating that the palatopharynx is closed. However, in the case of a VPI patient vocalizing, the switch is closed, indicating that the palatopharynx is not fully closed, thus allowing air to flow into the nasal passage. The system responses include the resonant and radiative effects of the resonant cavity on the vocal cord waves, which can be reflected in the spectrograms of vowels. Furthermore, the spectrograms of vowels change over time compared to the spectrum. In this work, the spectrograms used as the inputs to the network may allow the network to extract the distinguishable features between VPI and non-VPI vowels.

2.4.3. Difference Feature Extractor

Difference Feature Extractor (DFE) is the backbone network of CARS-Net. DFE is mainly used to extract deep features of the input image for subsequent classification. The DFE uses a Siamese-like network structure with two branching networks to extract the differences in the resonance response features contained in spectrograms of the two acoustic signals of the nasal channel and the oral channel.

The two branching networks of the DFE extract the deep features of the two input spectrograms of the nasal and oral cavities. The two branching networks in CARS-Net use a ResNet-18 architecture consisting of residual blocks. The residual block [47] is proposed to alleviate the problem of gradient disappearance due to increasing depth in deep convolutional neural networks by using jump connections.

In this paper, a cross-attention module (CA module) is designed to link the two branching networks. This means embedding a cross-attention block in the residual blocks in the two branching networks. The two input acoustic signals are correlated, and they can be expressed as two signals after obtaining different system responses for a homologous signal, as shown in (17). The traditional Siamese network structure has no connection between the two branching networks. The two branching networks share weights but process the two inputs independently without utilizing the correlation information between the two inputs. The CA module is implemented by coupling the intermediate feature maps of the two channels. This allows the originally independent branching networks to be linked in the feature extraction process, enhancing the model's ability to control the global information. Figure 5 shows the detailed structure of the proposed cross-attention block embedded in the residual block.

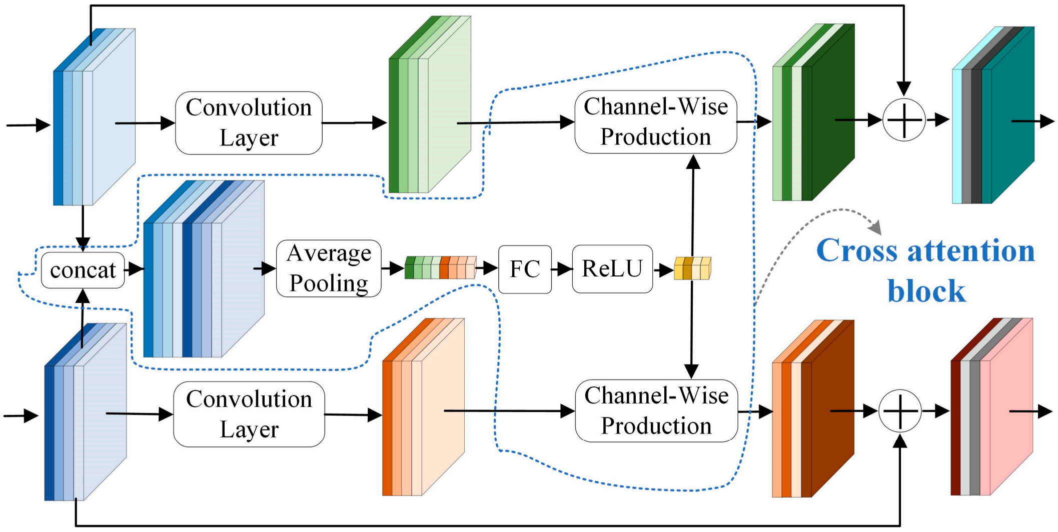

Figure 5. Cross-attention module.

The input feature maps of the *i*-th residual block in each of the two backbone networks are denoted as $F_{n\_i} \in R^{H_i \times W_i \times C_i}$ and $F_{o\_i} \in R^{H_i \times W_i \times C_i}$ respectively. The two input feature maps are concatenated in the channel dimension to obtain the fused feature maps, $F_{cct\_i} \in R^{H_i \times W_i \times 2C_i}$. Average pooling is performed on F_{cct_i} in the channel dimension. Pooling operations are used to refine the global information and reduce the number of operations [48].

Afterward, a fully connected layer operation is performed to output a weight vector of dimension C_i'. The nonlinearity of the fully connected layer can better fit the information of each channel of the feature maps. Then, a ReLU activation function is used to obtain the final weight vector $W_i \in R^{1\times1\times C_{i+1}}$.

The output feature maps of the feature extraction module ($F'_{n\_i} \in R^{H_{i+1}\times W_{i+1}\times C_{i+1}}$, $F'_{o\_i} \in R^{H_{i+1}\times W_{i+1}\times C_{i+1}}$) are numerically dot-multiplied with the weight vectors in the channel dimension to obtain the weighted feature maps ($M_{n\_i} \in R^{H_{i+1}\times W_{i+1}\times C_{i+1}}$, $M_{o\_i} \in R^{H_{i+1}\times W_{i+1}\times C_{i+1}}$).

Finally, the weighted feature maps are used to replace the original feature maps with the input feature maps in order to obtain the output feature maps of the *i*-th block, as shown in (14).

$$\begin{cases} F_{n\_i+1} = F_{n\_i} + M_{n\_i} \\ F_{o\_i+1} = F_{o\_i} + M_{o\_i} \end{cases} \tag{14}$$

where F_{n_i+1} is both the output of the *i*-th residual block and the input of the *i* + 1-th residual block.

2.4.4. VPI/Non-VPI Classifier at the Phoneme Level for Vowels

CARS-Net uses a fully connected layer for end-to-end VPI/non-VPI classification instead of threshold judgment, in contrast to the traditional Siamese network. The difference feature map obtained by the two branching networks is flattened into vectors, which are then fed sequentially into a fully connected layer and a softmax layer to achieve binary classification. The process is shown in (15).

$$output = softmax(FC(flatten(F_{n\_final} - F_{o\_final}\Big) \tag{15}$$

In the traditional Siamese network structure, the Euclidean distances of the final feature maps obtained from the two backbone networks are calculated for threshold classification, as shown in (16).

$$d = \Big|\Big|F_{n\_final} - F_{o\_final}\Big|\Big|_2 \tag{16}$$

where d represents the Euclidean distance of the final feature maps, and $||\cdot||_2$ is the L2 norm.

This classification method compresses the feature differences between the two channel signals into a single value, with an eye on the overall differences between the dual-channel inputs. This is equivalent to assigning the same weight to each local feature difference in the classification. In contrast, in the classification task of this work, the individual detailed differences in the two input spectrograms reflect different system response features. Different pronunciation system response features should also make different contributions to the classification. The fully connected layer is a nonlinear operation that assigns different classification weights to features at different positions in the difference feature map.

2.4.5. Loss Function

The loss function of CARS-Net consists of two parts, contrast entropy loss and cross-entropy loss, as shown in (17).

$$Loss = loss1 + loss2 \tag{17}$$

where $loss1$ is the cross-entropy loss for the VPI/non VPI classification results, and $loss1$ is the contrast entropy loss.

The cross-entropy loss captures the final classification accuracy, as shown in (18).

$$loss1 = -(y_i \log(P(\hat{y}_i)) + (1 - y_i)\log((1 - P(\hat{y}_i)))) \tag{18}$$

where $P(\hat{y}_i)$ is the predicted probability of existing VPI for the *i*-th sample.

Scholars [49] have designed the contrast entropy function as a loss function for network training in a traditional Siamese network based on the differences in the feature maps of the two branching networks. The contrast entropy loss provides direction for the overall feature extraction, as shown in (19).

$$loss2 = -\left(y_i d^2 + (1 - y_i)\max(\mathrm{margin} - d, 0)^2\right) \tag{19}$$

where the margin takes 2, and y_i is the label of the i-th sample.

3. Results and Analysis

3.1. Experiment Settings

For VPI/non-VPI classification at the phoneme level for consonants, two feature sets, *RPFD* and *RFD*, are proposed based on a power spectral density ratio sequence. The VPI/non-VPI classification model at the phoneme level for unvoiced consonants is then obtained by the SVM classifier.

For VPI/non-VPI classification at the phoneme level for vowels, a network structure, CARS-Net, is proposed. The parameters are initialized using the He initialization method [50] for training CARS-Net. The optimizer is selected as Adam [51]. Regarding the hyperparameter settings, the number of epochs is set to 100, the batch size is 64, and the learning rate is 0.001. The size of the spectrogram is 64 × 64.

The VPI/non-VPI classification experiment at the phoneme level for consonants and vowels uses 10-fold cross-validation. To decrease the error introduced by the detection results of individual consonants or vowels, a voting mechanism is used for the classification of the VPI for each patient/subject. The classification results of all phonemes participating in the experiment for a given subject are counted. A subject is classified as a VPI patient if more than half of the phonemes of the subject were identified as VPI speech.

3.2. VPI Detection Results at the Subject Level

Table 1 shows the VPI/non-VPI classification results at the subject level, represented by six classification evaluation metrics, namely, accuracy, precision, recall, F1-score, true negative rate (TNR), and false positive rate (FPR). Accuracy represents the ratio of the number of all correctly predicted VPI and non-VPI samples to the total number of samples in the dataset. Precision represents the proportion of all subjects predicted to be VPI patients who are actually VPI patients. Recall represents the proportion of VPI patients in the dataset who are correctly classified. The F1-score is the summed average of precision and recall, which combines the values of precision and recall and ranges from 0 to 1. The closer the value is to 1, the better the performance of the classification model.

Table 1. VPI detection results at the subject level (%).

Accuracy	Precision	Recall	F1-Score	TNR	FPR
98.52	97.80	100.00	98.89	95.65	4.35

As shown in Table 1, the accuracy of VPI detection at the subject level reached 98.52% based on the method proposed in this paper. These are the results for the binary classification of all VPI and non-VPI subjects in the dataset. The precision and recall were 97.8% and 100%, respectively. The F1-score reached 98.89%. It is shown that the proposed method can correctly detect all VPI patients in the dataset. In clinical diagnosis, misdiagnosis and omission can affect the treatment of VPI patients, leading to serious medical errors. A method with 0% missed detection and 2.2% misdiagnosis on this dataset might play a supporting role in clinical VPI diagnosis.

As shown in Table 1, the TNR and FPR are 95.65% and 4.35%, respectively. This means that 95.65% of the subjects in the dataset that are non-VPI controls were correctly predicted. In the clinic, false-positive diagnoses can lead to misdiagnosis and delay the treatment that the patient should receive. A low false positive rate is important in the clinical diagnosis of VPI.

In terms of the VPI/non-VPI classification results at the subject level, the voting method can improve the robustness of the automatic VPI/non-VPI detection system compared to using individual phoneme classification results. In the clinical diagnosis of VPI, speech therapists synthesize the auditory perception of a whole paragraph rather than a single phoneme. This paper uses a voting mechanism to perform automatic subject level VPI/non-VPI detection based on our proposed phoneme level detection algorithm. This method is consistent with clinical diagnosis.

3.3. VPI/Non-VPI Classification Results at the Phoneme Level for Consonants

This subsection discusses the VPI speech classification results for consonants. In the VPI/non-VPI classification method for consonants, two sets of features, RFD and $RPFD$, are extracted based on the power spectral density ratio sequence of the two-channel acoustic signals collected from the nasal and oral cavities, respectively. Table 2 shows the average VPI speech classification accuracy at the phoneme level for consonants based on $RPFD$, RFD, and combinations of the two sets of features with three different classifiers.

Table 2. VPI speech classification accuracy at the phoneme level for consonants (%).

Features Set	SVM	LDA	Adaboost
$RPFD$	76.19	76.81	78.82
RFD	84.81	84.74	80.10
$RPFD + RFD$	85.00	84.85	83.30

As shown in Table 2, the VPI speech classification accuracies of consonants based on $RPFD$ range from 76.19% to 78.82%. The classification accuracies of VPI speech based on RFD features range from 80.10% to 84.81%. RFD has more differentiation of classification than $RPFD$ for VPI speech and non-VPI speech) in this dataset.

Both $RPFD$ and RFD are features extracted on power spectral density ratio sequences of the nasal and oral channel acoustic signals. $RPFD$ is extracted to reflect the most prominent and abrupt frequencies of the acoustic signals of the nasal channel compared to the acoustic signals of the oral channel, while RFD is a reflection of the relative frequency distribution of the two channels. $RPFD$ is a reflection of specific frequency values, which are more affected by noise and less robust than RFD. As shown in the VPI speech classification results, the accuracy of the classification model based on RFD features is higher than that of the classification model based on $RPFD$ features.

As shown in Table 2, the VPI speech classification accuracy of the combination of $RPFD$ and RFD is 83.30–85.00%, which is higher than that of single-set features. The two sets of features, $RPFD$ and RFD, complement each other for VPI and non-VPI speech binary classification model descriptions.

3.4. VPI/Non-VPI Classification Results at the Phoneme Level for Different Consonants

This subsection discusses the results of VPI speech classification for different consonants to investigate the effect of articulatory processes and articulatory organs on the proposed VPI consonant classification method.

Three types of consonants with different articulatory processes are involved in this experiment: the aspirated plosives (/p/, /t/, /k/), aspirated affricates (/q/, /c/), and fricatives (/h/, /x/, /sh/, /f/). These consonants are classified by articulatory organ into bilabial (/p/), alveolar (/t/), velar (/k/, /h/), front palatal (/q/, /x/), blade alveola (/c/), retroflexes (/sh/), and labiodental (/f/) as shown in Figure 6.

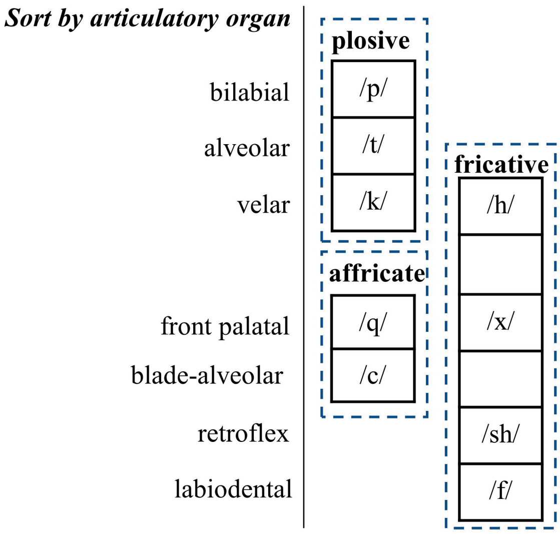

Figure 6. Classification of the unvoiced consonants by the articulation organs.

Table 3 shows the VPI speech classification results for different consonants using the method proposed in this paper.

Table 3. VPI speech classification accuracy at the phoneme level for different consonants (%).

Categories of Different Initial Consonants			**Classifiers**		
			SVM	**LDA**	**Adaboost**
bilabial	plosive	/p/	80.77	80.77	81.54
alveolar	plosive	/t/	77.69	77.69	80.00
velar	plosive	/k/	80.94	80.14	73.13
	fricative	/h/	83.33	83.33	81.05
front palatal	affricate	/q/	89.37	91.00	90.97
	fricative	/x/	91.32	92.12	88.12
blade alveolar	affricate	/c/	87.32	86.58	87.32
retroflex	fricative	/sh/	84.17	84.17	85.00
labiodental	fricative	/f/	90.11	87.83	82.56

As shown in Table 3, the VPI speech classification accuracy for plosives (/p/, /t/, /k/) ranges from 73.13% to 81.54%; the VPI consonant classification accuracy for affricates (/q/, /c/) ranges from 86.58% to 91.00%; and the accuracy for fricatives (/h/, /x/, /sh/, /f/) ranges from 81.05% to 92.12%. The accuracy of VPI consonant classification for plosives is lower compared to that for affricates and fricatives.

VPI leads to a partial overflow of air to the nasal cavity and a decrease in oral airflow. After the deblocking process of affricates and fricatives, the airflow passes through small gaps shaped by the articulatory organs [52], and the plosives involve a complete closure and lack the air friction process [52].Therefore, the reduction of oral airflow has less effect on the pronunciation of plosives than affricates and fricatives.

As shown in Table 3, for the same articulatory organ (/q/, /x/), the fricative-based VPI speech classification accuracy (/x/) is 92.12%, which is higher than the affricate-based VPI speech classification accuracy (/q/) of 91.00%. In terms of vocalization duration, that of fricatives is the longest, that of affricates is the next shortest, and that of plosives is the shortest [52,53]. As the duration of the vocalization increases, the airflow continues to spill into the nasal cavity due to VPI. This results in a more pronounced decrease in air pressure in the oral cavity. At this time, the change in the relative spectrum distribution of radiated signals from the nasal and oral cavities is greater. Therefore, the non-VPI and VPI speech signals are more distinguishable according to the fricative-based features proposed in this paper.

As shown by the VPI speech classification results from the perspective of different articulatory organs in Table 3, the front palatal-based (/q/, /x/) classification accuracy is optimal with a maximum of over 91%. The highest blade alveolar-based, retroflex-based, and labiodental-based classification accuracies are above 85%. The velar (/h/) accuracy obtained the lowest accuracy, only 81.05%, among all the affricates and fricatives.

The front palatals (/q/, /x/) are pronounced with the front of the tongue against or near the front of the hard palate, where the airflow is obstructed and then formed. The location of articulatory deblocking for the front palatals (/q/, /x/) is at the hard palate. VPI causes a gap in the plane where the palatopharynx and hard palate are supposed to form, causing it to further interfere with the deblocked articulation process of the hard palate. This makes front palatals (/q/, /x/) more distinguishable in VPI/non-VPI classification at the phoneme level for consonants.

The soft palate is the deblocking site for the velar (/h/). The soft palate is located closer to the palatopharynx than the rest of the affricates and fricatives. VPI results in less impact on the reliance on soft palate position to deblock vocalization when airflow is shunted in the palatopharynx. This results in lower differentiation between the non-VPI and VPI pronunciations of the velar (/h/).

3.5. Effect of the Parameter of RFD on VPI/Non-VPI Classification for Unvoiced Consonants

N_h is the number of subbands that are linearly divided from the whole frequency band. The different values of N_h divide the frequency band into different numbers of frequency bands in the *RFD*. This subsection explores the effect of different values of N_h on the VPI/non-VPI classification for consonants.

As shown in Table 4, the VPI/non-VPI speech classification accuracies are 84.81%, 84.32%, and 84.80% when N_h is taken as 4, 8, and 16, respectively. The increase in N_h does not have an improvement on the accuracy of VPI/non-VPI speech classification at the phoneme level for consonants.

Table 4. VPI/non-VPI classification results at the phoneme level for different vowels (%).

	N_h=4	N_h=8	N_h=16
Accuracy	84.81	84.32	84.80

The consonant pronunciation of VPI patients presents stronger nasal radiation acoustic signals than that of non-VPI controls. It produces a change in the percentage of the concentrated frequency band of the oral radiation acoustic signals in the whole frequency band. The frequency of most unvoiced consonants in Mandarin radiated from the oral cavity are higher than 4000 Hz [54]. When N_h is taken as 4, a frequency band has a frequency range of 4134–5502.5 Hz, which is almost coincident with the concentrated frequency bands of the consonants radiated from the oral cavity. Therefore, N_h takes an empirical value of 4 in this paper.

3.6. Analysis of VPI Speech Classification Results Based on Different Vowels

When nonnasalized vowels are produced by VPI patients, the vocal cord wave spills into the nasal cavity, which does not occur without VPI. In this paper, CARS-Net is proposed to perform automatic VPI/non-VPI speech classification for vowels. It captures the different feature maps of the spectrum of the acoustic signals of the nasal and oral channels to distinguish vowels produced by patients with VPI and without VPI. Table 5 shows the results of automatic VPI/non-VPI speech classification for different vowels.

Table 5. VPI/non-VPI classification result at the phoneme level for different vowels (%).

	Accuracy	Precision	Recall	F1-Score
/a/	89.73	92.06	92.92	92.19
/e/	90.36	93.92	93.55	93.42
/i/	93.72	95.42	96.11	95.45
/u/	93.42	95.40	95.25	95.10

From the VPI/non-VPI speech detection results of the four vowels (/a/, /i/, /e/, and /u/) shown in Table 5, the classification accuracies of both /i/ and /u/ exceed 93%, while the classification accuracies of /a/ and /e/ are approximately 90%. The recall of 92.92% and 93.55% for /a/ and /e/ is also lower than the average recall of over 95% for /i/ and /u/. This indicates that the model is more sensitive in classifying the VPI/non-VPI speech of /i/ and /u/ than /a/ and /e/. This is consistent with the conclusions reached in the works [36,37], which does not provide an explanation for this phenomenon. In this paper, it is proposed that this phenomenon is related to the level of tongue position when pronouncing vowels.

The shapes of the tongue and lip constitute different oral resonator shapes, thus producing different vowel sounds. The position of the tongue describes the vertical distance between the upper surface of the tongue and the palate [55]. According to the position of the tongue, /a/ is a low vowel, /e/ is a semihigh vowel, and /i/ and /u/ are high vowels. The higher the tongue position, the closer the tongue is to the palate, and the narrower the airflow passage between the tongue and the palate. A narrow passage impedes the passage of airflow more than a wide passage.

If velopharyngeal function is normal, the airflow only moves toward the oral cavity when nonnasalized sounds are produced. In the presence of VPI, the airflow is directed to both the nasal and oral cavities. When producing higher lingual vowels, the narrower airflow passage between the tongue and the palate may force more airflow toward the nasal cavity. This results in more vocal cord waves flowing into the nasal cavity and fewer into the oral cavity. The difference between the acoustic signals of the nasal and oral channels is more distinguishable between VPI high vowels and non-VPI high vowels. This may explain why the CARS-Net proposed in this paper is more sensitive to high vowels (/i/, /u/) in automatic VPI/non-VPI speech classification for vowel tasks

3.7. Effectiveness of the Cross-Attention Module

To verify the effectiveness of the proposed cross-attention module in this paper, ablation experiments are conducted in this section. Table 6 shows the comparison of the model prediction results after training the network without and with the CA module using the same hyperparameters.

Table 6. VPI/non-VPI speech classification for vowels with CA-module or not (%).

Vowel	Structure	Accuracy	Precision	Recall	F1-Score
/a/	No_CA [1]	80.24	81.67	90.69	85.71
	CARS-Net	89.73	92.06	92.92	92.19
/e/	No_CA [1]	88.00	90.59	92.13	91.08
	CARS-Net	90.93	93.84	93.52	93.36
/i/	No_CA [1]	89.25	95.22	89.44	91.76
	CARS-Net	93.71	95.42	96.09	95.44
/u/	No_CA [1]	93.13	95.09	95.22	94.89
	CARS-Net	93.42	95.39	95.27	95.11

[1] CRAS-Net without CA module.

As shown in Table 6, for each different vowel, the model prediction results of the network model containing the CA-module outperformed the training model without the CA-module in all four metrics (accuracy, precision, recall, and F1-score), showing the effectiveness of the proposed CA-module for VPI recognition in this paper. This shows that the CA-module can improve the ability of the traditional Siamese network to extract the correlation features of the two channel acoustic signals for VPI/non-VPI speech classification.

The accuracy of the network model with the CA module improved by 9.49%, 2.93%, 4.46%, and 1.04% on /a/, /e/, /i/, and /u/, respectively, compared with the network model without the CA module. On the one hand, the CA module improved the VPI/non-VPI speech classification results for /a/ most significantly, which is the vowel with the lowest sensitivity in VPI/non-VPI speech classification among the four vowels involved in the experiment. On the other hand, although the network models containing the CA module show different improvements in VPI/non-VPI speech classification based on all four vowels, the VPI speech recognition accuracies of /a/ and /e/ are still lower than those of /i/ and /u/ when using the CA module. The CA module improves the overall recognition effectiveness of the network for VPI speech without changing the relative effectiveness for different vowels. This further indicates that the VPI/non-VPI speech classification method for vowels proposed in this paper has higher discriminative power for /i/ and /u/ than for /a/ and /e/.

3.8. Validation of the Loss Function

The loss function used in this paper contains two parts, the cross-entropy loss ($loss1$), for binary classification results, and the contrast entropy loss ($loss2$) between the feature maps of two branching networks. To verify the effectiveness of the loss function, the classification results of using only the cross-entropy loss for network training are compared with the results of using the loss function $Loss$ ($loss1 + loss2$). The two network models are trained with the same hyperparameter settings. The results are shown in Table 7.

Table 7. VPI/non-VPI speech classification for vowels results with loss2 or not (%).

Vowel	Loss	Accuracy	Precision	Recall	F1-Score
/a/	$loss1$	71.53	86.08	74.17	76.15
	$loss1 + loss2$	89.73	92.06	92.92	92.19
/e/	$loss1$	83.66	89.36	86.80	87.34
	$loss1 + loss2$	90.93	93.84	93.52	93.36
/i/	$loss1$	90.79	94.16	92.43	93.20
	$loss1 + loss2$	93.71	95.42	96.09	95.44
/u/	$loss1$	88.91	93.69	89.49	90.26
	$loss1 + loss2$	93.42	95.39	95.27	95.11

As shown, the overall effectiveness of the network model training without the contrast entropy loss is inferior to that of the network model trained by the loss function with the contrast entropy loss. This indicates that the cross-entropy loss provides an optimization direction for difference feature extraction between VPI patients and non-VPI controls.

3.9. Comparison with the State of Art

To verify the validity of the proposed network, this subsection describes experiments with three existing classical classification networks with single-channel inputs, namely VGG16 [56], AlexNet [57], and ResNet18 [47]. The spectrograms of the oral and nasal channels are concatenated as inputs to the single-channel network. Table 8 shows the average results of the above three networks, as well as CARS-Net, on the /i/ and /u/.

Table 8. Comparison experiment results of VPI/non-VPI speech classification for vowels (%).

	Accuracy	**Precision**	**Recall**	**F1-Score**
VGG16	83.21	84.47	94.23	87.95
AlexNet	90.96	93.36	94.09	93.25
Resnet18	92.60	94.74	94.79	94.53
CARS-Net	93.57	95.41	95.67	95.27

The experimental results in Table 8 show that ResNet18 has higher accuracy than VGG16 and AlexNet. The CARS-Net using ResNet18 as the branching network proposed in this paper improved in accuracy, precision, recall, and F1-score compared with ResNet18 using a single channel. This indicates the effectiveness of the difference in deep features between the spectrograms of nasal- and oral-channel acoustic signals for VPI/non-VPI vowel classification at the phoneme level.

4. Conclusions

The rise of artificial intelligence technology has brought new solutions to the scarcity of healthcare resources. The scarcity of speech therapists has prevented large-scale assurance of diagnosis and speech disorder assessment for patients with VPI. This paper presents an automatic system for VPI detection at the subject level. Regarding VPI/non-VPI classification for unvoiced consonants, relative prominent feature description and relative feature distribution features are shown to be effective. Furthermore, the effect of unvoiced consonants with different articulatory organs on the production of VPI symptoms is explored. Regarding vowel classification, the cross-attention module embedded in CARS-Net has been proven to be effective in the VPI/non-VPI classification task. Furthermore, the sensitivity of CARS-Net for vowels with different tongue positions on VPI classification is explored. CARS-Net is more sensitive to high vowels than to vowels of lower tongue position for VPI/non-VPI classification. The experimental results obtained by voting on the phoneme level results also demonstrate the effectiveness of the proposed system for VPI recognition at the subject level. The performance of the system makes screening for VPI and assessment for speech disorders possible even in a global shortage of speech therapists.

5. Future Work

An automatic ancillary diagnostic approach to speech-based VPI is proposed in the work. It is the achievement of qualitative diagnosis for VPI patients. In the clinical treatment of VPI patients, the assessment of the severity of VPI is also important. This not only helps in developing treatment plans for VPI patients, but also provides an objective basis for the recovery process. In the future work, we will explore the feasibility of assessing severity of VPI patients based on speech data.

Author Contributions: Conceptualization, Y.Z. and L.H.; methodology, Y.Z. and L.H.; software, Y.Z. and J.Z.; validation, L.H., Y.Z. and W.L.; formal analysis, J.Z. and W.L.; investigation, all authors; resources, H.Y. and J.Z.; data curation, H.Y.; writing—original draft preparation, Y.Z. and L.H.; writing—review and editing, all authors; visualization, W.L.; supervision, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61503264, and the Exploration and the Research Project of West China Hospital of Stomatology, Sichuan University, grant number LCYJ-2022-YF-3.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of West China Hospital of Stomatology, Sichuan University (protocol code WCHSIRB-D-2021-205, and the date of approval is 3 June 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The datasets presented in this article include identifiable sound information that is privacy-preserving and must be approved by the West China Hospital of Stomatology, Sichuan University. Requests to access the datasets should be directed to L.H., ling.he@scu.edu.cn.

Acknowledgments: We thank all the participants for their contributions and the support of experimental data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Evans, A.; Ackermann, B.; Driscoll, T. Functional anatomy of the soft palate applied to wind playing. *Med. Probl. Perform. Artist.* **2010**, *25*, 183–189. [CrossRef]
2. Lam, D.J.; Chiu, L.L.; Sie, K.C.; Perkins, J.A. Impact of cleft width in clefts of secondary palate on the risk of velopharyngeal insufficiency. *Arch. Facial Plast. Surg.* **2012**, *14*, 360–364. [CrossRef] [PubMed]
3. Skirko, J.R.; Weaver, E.M.; Perkins, J.; Kinter, S.; Sie, K.C. Modification and evaluation of a velopharyngeal insufficiency quality-of-life instrument. *Arch. Otolaryngol.* **2012**, *138*, 929–935. [CrossRef]
4. Lynch, C.A.; Rule, D.W.; Klaben, B.; Oren, L.; Friedman, A.D.; Khosla, S.M.; Howell, R.J. Surgical Treatment of Acquired Velopharyngeal Insufficiency in Adults with Dysphagia and Dysphonia. *J. Voice*, 2022; *in press*. [CrossRef] [PubMed]
5. Conley, S.F.; Gosain, A.K.; Marks, S.M.; Larson, D.L. Identification and assessment of velopharyngeal inadequacy. *Am. J. Otolaryng.* **1997**, *18*, 38–46. [CrossRef]
6. Young, K.; Sweeney, T.; Vos, R.R.; Mehendale, F.; Daffern, H. Evaluation of noise excitation as a method for detection of hypernasality. *Appl. Acoust.* **2022**, *190*, 108639. [CrossRef]
7. Silva-Rojas, A.; Ysunza, A.; Diaz-Torres, D.; Bardales-Lazcano, M.; Pamplona, M.C. Velopharyngeal insufficiency as the initial manifestation of a Myotonic Dystrophy type 1: A case report. *Int. J. Pediatr. Otorhi. Extra* **2012**, *7*, 190–192. [CrossRef]
8. D'Antonio, L.L.; Snyder, L.S.; Samadani, S. Tonsillectomy in children with or at risk for velopharyngeal insufficiency: Effects on speech. *Otolaryng. Head Neck* **1996**, *115*, 319–323. [CrossRef]
9. Saunders, N.; Hartley, B.; Sell, D.; Sommerlad, B. Velopharyngeal insufficiency following adenoidectomy. *Clin. Otolaryngol.* **2004**, *29*, 686–688. [CrossRef]
10. Raol, N.; Hartnick, C.J. Anatomy and physiology of velopharyngeal closure and insufficiency. In *Surgery for Pediatric Velopharyngeal Insufficiency*; Karger Medical and Scientific Pulishers: Basel, Switzerland, 2015; Volume 76, pp. 1–6. [CrossRef]
11. Hardin-Jones, M.; Karnell, M.; Peterson-Falzone, S.; Trost-Cardamone, S. *The Clinician's Guide to Treating Cleft Palate Speech*; Elsevier: St. Louis, MO, USA, 2017; pp. 20–31.
12. Sell, D.; Grunwell, P.; Mildinhall, S.; Murphy, T.; Cornish, T.A.; Bearn, D.; Shaw, W.C.; Murray, J.J.; Williams, A.C.; Sandy, J.R. Cleft lip and palate care in the United Kingdom—The Clinical Standards Advisory Group (CSAG) Study. Part 3: Speech outcomes. *Cleft Palate-Cran. J.* **2001**, *38*, 30–37. [CrossRef]
13. Kummer, A.W. Speech therapy for errors secondary to cleft palate and velopharyngeal dysfunction. *Semin. Speech Lang.* **2011**, *32*, 191–198. [CrossRef] [PubMed]
14. Shadi, M.S.; Hegazi, M.A.; Ghandour, H.H.; Othman, N.M.; Khaled, D.M. Patterns of velopharyngeal closure during speech in individuals with normal habitual resonance: A nasoendoscopic analysis. *Auris Nasus Larynx* **2022**, *49*, 995–1002. [CrossRef] [PubMed]
15. Lam, D.J.; Starr, J.R.; Perkins, J.A.; Lewis, C.W.; Eblen, L.E.; Dunlap, J.; Sie, K.C. A comparison of nasendoscopy and multiview videofluoroscopy in assessing velopharyngeal insufficiency. *Otolaryng. Head Neck* **2006**, *134*, 394–402. [CrossRef] [PubMed]
16. Raol, N.; Caruso, P.; Hartnick, C.J. Use of imaging to evaluate course of the carotid artery in surgery for velopharyngeal insufficiency. *Ann. Otol. Rhinol. Laryngol.* **2015**, *124*, 261–265. [CrossRef] [PubMed]
17. Kobayashi, Y.; Kobayashi, M.; Kanamori, D.; Fujii, N.; Kataoka, Y.; Satoh, K.; Sano, Y.; Yoshioka, S.; Tateya, I.; Toyama, H. Evaluation of Velopharyngeal Closure Function With 4-Dimensional Computed Tomography and Assessment of Radiation Exposure in Pediatric Patients: A Cross-Sectional Study. *Cleft Palate-Cran. J.* **2022**, *59*, 141–148. [CrossRef]
18. Kobayashi, Y.; Kanamori, D.; Fujii, N.; Kataoka, Y.; Hirai, E.; Yoshioka, S.; Satoh, K.; Toyama, H.; Naito, K.; Matsuo, K. Velopharyngeal closure analysis using four-dimensional computed tomography: A pilot study of healthy volunteers and adult patients with cleft palate. *BMC Med. Imaging* **2019**, *19*, 1–8. [CrossRef]
19. Atik, B.; Bekerecioglu, M.; Tan, O.; Etlik, O.; Davran, R.; Arslan, H. Evaluation of dynamic magnetic resonance imaging in assessing velopharyngeal insufficiency during phonation. *J. Craniofac. Surg.* **2008**, *19*, 566–572. [CrossRef]
20. Denegri, M.A.; Silva, P.P.; Pegoraro-Krook, M.I.; Ozawa, T.O.; Yaedu, R.Y.F.; Dutka, J.D.C.R. Cephalometric predictors of hypernasality and nasal air emission. *J. Appl. Oral Sci.* **2021**, *29*, e20210320. [CrossRef]
21. Lou, Q.; Liu, T.; Wang, X.; Wu, D.; Wang, G.; Chen, Y.; Wan, T. An Observational Study to Evaluate Association between Velopharyngeal Anatomy and Speech Outcomes in Adult Patients with Severe Velopharyngeal Insufficiency. *J. Craniofac. Surg.* **2021**, *32*, 2753–2757. [CrossRef]
22. Van Lierde, K.M.; Claeys, S.; De Bodt, M.; Van Cauwenberge, P. Outcome of laryngeal and velopharyngeal biofeedback treatment in children and young adults: A pilot study. *J. Voice* **2004**, *18*, 97–106. [CrossRef]

23. Bettens, K.; Wuyts, F.L.; Van Lierde, K.M. Instrumental assessment of velopharyngeal function and resonance: A review. *J. Commin. Disord.* **2014**, *52*, 170–183. [CrossRef] [PubMed]
24. De Boer, G.; Bressmann, T. Comparison of nasalance scores obtained with the nasometers 6200 and 6450. *Cleft Palate-Cran. J.* **2014**, *51*, 90–97. [CrossRef] [PubMed]
25. Karnell, M.P. Instrumental assessment of velopharyngeal closure for speech. *Semin. Speech Lang.* **2011**, *32*, 168–178. [CrossRef] [PubMed]
26. de Stadler, M.; Hersh, C. Nasometry, videofluoroscopy, and the speech pathologist's evaluation and treatment. In *Surgery for Pediatric Velopharyngeal Insufficiency*; Karger Medical and Scientific Pulishers: Basel, Switzerland, 2015; Volume 76, pp. 7–17. [CrossRef]
27. Kataoka, R.; Warren, D.W.; Zajac, D.J.; Mayo, R.; Lutz, R.W. The relationship between spectral characteristics and perceived hypernasality in children. *J. Acoust. Soc. Am.* **2001**, *109*, 2181–2189. [CrossRef]
28. Maeda, S. Acoustics of vowel nasalization and articulatory shifts in French nasal vowels. In *Nasals, Nasalization, and the Velum*; Elsevier: Amsterdam, The Netherlands, 1993; Volume 5, pp. 147–167.
29. Vijayalakshmi, P.; RamasubbaReddy, M. The analysis on band-limited hypernasal speech using group delay based formant extraction technique. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
30. Dubey, A.K.; Prasanna, S.M.; Dandapat, S. Zero time windowing analysis of hypernasality in speech of cleft lip and palate children. In Proceedings of the 2016 Twenty Second National Conference on Communication (NCC), Guwahati, India, 4–6 March 2016; pp. 1–6.
31. Vijayalakshmi, P.; Reddy, M.R.; O'Shaughnessy, D. Acoustic analysis and detection of hypernasality using a group delay function. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 621–629. [CrossRef]
32. Vijayalakshmi, P.; Nagarajan, T.; Ra, V.J. Selective pole modification-based technique for the analysis and detection of hypernasality. In Proceedings of the TENCON 2009–2009 IEEE Region 10 Conference, Singapore, 23–26 November 2009; pp. 1–5.
33. Dubey, A.K.; Tripathi, A.; Prasanna, S.; Dandapat, S. Detection of hypernasality based on vowel space area. *J. Acoust. Soc. Am.* **2018**, *143*, EL412–EL417. [CrossRef]
34. Dubey, A.K.; Prasanna, S.R.; Dandapat, S. Detection and assessment of hypernasality in repaired cleft palate speech using vocal tract and residual features. *J. Acoust. Soc. Am.* **2019**, *146*, 4211–4223. [CrossRef]
35. Dubey, A.K.; Prasanna SR, M.; Dandapat, S. Sinusoidal model-based hypernasality detection in cleft palate speech using CVCV sequence. *Speech Commun.* **2020**, *124*, 1–12. [CrossRef]
36. Wang, X.; Yang, S.; Tang, M.; Yin, H.; Huang, H.; He, L. HypernasalityNet: Deep recurrent neural network for automatic hypernasality detection. *Int. J. Med. Inform.* **2019**, *129*, 1–12. [CrossRef]
37. Wang, X.; Tang, M.; Yang, S.; Yin, H.; Huang, H.; He, L. Automatic hypernasality detection in cleft palate speech using cnn. *Circ. Syst. Signal. Pract.* **2019**, *38*, 3521–3547. [CrossRef]
38. Mathad, V.C.; Scherer, N.; Chapman, K.; Liss, J.; Berisha, V. An Attention Model for Hypernasality Prediction in Children with Cleft Palate. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7248–7252.
39. Mathad, V.C.; Scherer, N.; Chapman, K.; Liss, J.M.; Berisha, V. A deep learning algorithm for objective assessment of hypernasality in children with cleft palate. *IEEE Trans. Biomed. Eng.* **2021**, *68*, 2986–2996. [CrossRef] [PubMed]
40. Song, K.; Wan, T.; Wang, B.; Jiang, H.; Qiu, L.; Xu, J.; Jiang, L.; Lou, Q.; Yang, Y.; Li, D.; et al. Improving Hypernasality Estimation with Automatic Speech Recognition in Cleft Palate Speech. *arXiv* **2022**, arXiv:2208.05122. [CrossRef]
41. Kuehn, D.P.; Henne, L.J. Speech evaluation and treatment for patients with cleft palate. *Am. J. Speech-Lang. Pat.* **2003**, *12*, 103–109. [CrossRef] [PubMed]
42. Wang, J.; Li, Y. Research on Children's Mandarin Chinese Voiceless Consonant Airflow. In Proceedings of the 2019 3rd International Conference on Art Design, Language and Humanities (ADLH 2019), Moscow, Russia, 4–5 October 2019; pp. 210–216.
43. Kalita, S.; Sudro, P.N.; Prasanna, S.M.; Dandapat, S. Nasal Air Emission in Sibilant Fricatives of Cleft Lip and Palate Speech. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 4544–4548.
44. Kummer, A.W.; Lee, L. Evaluation and treatment of resonance disorders. *Lang. Speech Hear. Ser.* **1996**, *27*, 271–281. [CrossRef]
45. Barkana, B.D. Detection of the Parameters of Hypernasality. In Proceedings of the 2009 Sixth International Conference on Information Technology: New Generations, Las Vegas, NV, USA, 27–29 April 2009; pp. 1262–1264.
46. Rong, P.; Kuehn, D.P. The effect of oral articulation on the acoustic characteristics of nasalized vowels. *J. Acoust. Soc. Am.* **2010**, *127*, 2543–2553. [CrossRef]
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Liu, J.-J.; Hou, Q.; Liu, Z.-A.; Cheng, M.-M. Poolnet+: Exploring the potential of pooling for salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 887–904. [CrossRef]
49. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; pp. 1735–1742.

50. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
51. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. [CrossRef]
52. Alwan, A.; Jiang, J.; Chen, W. Perception of Place of Articulation for Plosives and Fricatives in Noise. *Speech Commun.* **2011**, *53*, 195–209. [CrossRef]
53. Tam, H.C. Common Pronunciation Problems of Vietnamese Learners of English. *VNU J. Fore.* **2005**, *21*, 35–46.
54. Wu, L.; Xiao, K.; Wang, S.; Wan, M. Acoustic influence of the neck tissue on Mandarin voiceless consonant production of electrolaryngeal speech. *Speech Commun.* **2017**, *87*, 31–39. [CrossRef]
55. Bourdiol, P.; Mishellany-Dutour, A.; Peyron, M.A.; Woda, A. Tongue-mandible coupling movements during saliva swallowing. *J. Oral rehabil.* **2014**, *41*, 199–205. [CrossRef] [PubMed]
56. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [CrossRef]
57. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.