applied sciences

MDPI

Article

Analysis of Phonetic Segments of Oesophageal Speech in People Following Total Laryngectomy

Krzysztof Tyburek, Dariusz Mikołajewski and Izabela Rojek *

Faculty of Computer Science, Kazimierz Wielki University, 85-064 Bydgoszcz, Poland
* Correspondence: izabela.rojek@ukw.edu.pl

Featured Application: Semi-automatic or automatic systems supporting rehabilitation of laryngectomised patients by improving the quality of oesophageal speech.

Abstract: This paper presents an approach to extraction techniques for speaker recognition following total laryngectomy surgery. The aim of the research was to develop a pattern of physical features describing the oesophageal speech in people after experiencing laryngeal cancer. Research results may support the speech rehabilitation of laryngectomised patients by improving the quality of oesophageal speech. The main goal of the research was to isolate the physical features of oesophageal speech and to compare their values with the descriptors of physiological speech. Words (in Polish) used during speech rehabilitation were analyzed. Each of these words was divided into phonetic segments from which the physical features of speech were extracted. The values of the acquired speech descriptors were then used to create a vector of the physical features of oesophageal speech. A set of these features will determine a model that should allow us to recognize whether the speech-rehabilitation process is proceeding correctly and also provide a selection of bespoke procedures that we could introduce to each patient. This research is a continuation of the analysis of oesophageal speech published previously. This time, the effectiveness of parameterization was tested using methodologies for analyzing the phonetic segments of each word.

Keywords: science; medical informatics; acoustic analysis; laryngectomy; phonetic segments; speech signal; vocal rehabilitation; voice pathology

Citation: Tyburek, K.; Mikołajewski, D.; Rojek, I. Analysis of Phonetic Segments of Oesophageal Speech in People Following Total Laryngectomy. *Appl. Sci.* **2023**, *13*, 4995. https://doi.org/10.3390/app13084995

Academic Editors: Chih-Ching Huang, Mukul Shirvaikar and Chung Hyun Goh

Received: 16 February 2023
Revised: 29 March 2023
Accepted: 13 April 2023
Published: 16 April 2023

1. Introduction

According to the Head Office of the National Health Fund, Department of Analyses and Innovation [1], laryngeal cancer is the most common malignant tumour among head and neck cancers. In the case of a very advanced neoplastic disease, a laryngectomy is necessary [2–5]. The consequence of this surgical procedure is the loss of the complete ability to communicate with the physiological voice. However, a patient deprived of a larynx has a chance to learn substitute speech during phoniatric rehabilitation. Therefore, the analysis of oesophageal speech is a very important research issue. Control of the course of the rehabilitation process should be supported by an observation of the physical features of speech. For this purpose, the values of speech descriptors of sick and healthy people are extracted in order to analyze their numerical differences. Professional methods of Signac processing and analysis, in particular acoustic methods, provide several options for assessing the quality of the speech signal, enabling multilateral analysis. Appropriately applied methods of speech analysis allow for the definition of a set of features that will determine the physical pattern of oesophageal speech [6,7]. In this paper, an attempt was made to analyze the voices of patients after total removal of the larynx (laryngectomy). The results of the tests will support the process of speech rehabilitation in order to increase the quality and reduce the time for complete rehabilitation. Speech rehabilitation of people following total laryngectomy can be carried out using three options [6,7]:

1. Oesophageal speech: This is a kind of substitute speech following laryngectomy. After a laryngectomy, the folds of the esophageal mucosa may act as a sound source of sound. This is the so-called pseudoglottis in the oesophagus. During oesophageal speech, it is necessary to swallow small amounts of air, which then comes back up via "burping". The column of swallowed air causes the oesophagus to vibrate and generate sound, which is modified by the tongue and lips to form words. The advantages of oesophageal speech is are:
   - Non-surgical method;
   - Hands-free talking;
   - Closest to physiological speech;
   - No need to implement a foreign body.

   Disadvantages of oesophageal speech are:
   - Learning takes a lot of time and must be intensive;
   - Not all people are able to master this method well;
   - Speech may be incomprehensible;
   - Speaking in short sentences and at a slower pace—having to swallow air while speaking.
2. Speaking with a voice prosthesis provides the most natural-sounding and easiest-to-understand voice. The prosthesis is placed between the oesophagus and the trachea during a total laryngectomy procedure. The prosthesis has a one-way valve that opens during speak and closes during breathing and eating. When speaking, it is necessary to close the valve with a finger. There are many models of voice prostheses, such as Provox. The advantages of voice prostheses are:
   - Ability to speak immediately after a laryngectomy;
   - Greater speech efficiency (no need to swallow air);
   - Clearer speech.

   The disadvantages of voice prostheses are:
   - Need to implant a foreign body, which may result in tinea or infections;
   - Periodic replacement;
   - Occurrence of leaks around the prosthesis;
   - Spontaneous prolapse of the prosthesis;
   - Appearance of inflammation.
3. Electrolarynx speech—this is an electrolarynx device. This method requires the use of a hand-held device which, when applied to the neck, generates vibrations that are then shaped by the tongue and mouth into speech. The voice produced by this method sounds very artificial and the modulation depends on the device used. Advantages:
   - Easy- to- master speech;
   - Non-surgical method.

   Disadvantages:
   - The artificial sound of speech;
   - The need to wear the device;
   - The need to use the hand when speaking;
   - Periodic service of the device required.

A very common opinion among laryngectomized people is that oesophageal speech is the real one, in the natural sense—in contrast to the use of speech prosthesis [8–10]. The analysis of oesophageal speech can be carried out by examining the spoken vowels, consonants, whole words, or individual fragments of these words separately. In [2], the authors focused on the parameterization of oesophageal speech by examining the pronounced vowel "i". Patients uttered the same vowels repeatedly over a period of time. In the literature, different languages are considered when examining oesophageal speech. This is due to the fact that each language has inherently different characteristics that have a direct

impact on the learning of oesophageal speech. This results from the particular articulation of consonants, vowels and individual words spoken in a particular language. In paper [4], the authors focused on the lexical tones of the Taiwanese language. This analysis took place in the time–frequency space of the signal. This research focused on the observation of changes in the fundamental frequency (F0), the slope of the F0 contour, and the duration and the amplitude of the vowels of parts of syllables containing seven Taiwanese tones. The study involved seven people undergoing speech rehabilitation after total laryngectomy [4]. In this paper, the acoustic analysis revealed no significant effects of the linguistic level on the acoustic parameters except for duration. This result seems justified due to the use of a small set of speech-signal physical characteristics of the speech signal. The authors in [8] attempted to improve the voice quality of oesophageal speech. In this case, working with time-, spectrum- and cepstrum-function descriptors aimed at improving the quality of oesophageal speech and thus increasing the effectiveness of communication. Available publications on the analysis of oesophageal speech prove that this analysis should take place in the time–frequency space and the cepstrum. In addition, the parameterization should take into account the language in which the test words are spoken. The study of oesophageal speech was also undertaken in [7]. The results of this analysis showed that the feature vectors obtained from the time domain, frequency domain, cepstrum and mel coefficients allow for precise parameterization of oesophageal speech. Audio features (including speech features) are extracted directly from the samples of the audio signal. Typical examples are the short-term energy and short-term zero-crossing rate. Such features offer a simple way to analyze audio signals, although it is usually necessary to combine them with more sophisticated frequency-domain features. Two approaches are possible:

- Based on the so-called signal macrostructure—calculations are performed in time segments after initial segmentation, the obtained parameters are the amplitude and rate of change;
- Based on the so-called the signal microstructure, i.e., the time course, analyzing the zero-crossing rate of the speech signal. This leads to obtaining two types of parameters: the density of zero crossings and the distribution of time intervals.

The aim of the authors of this paper was to parameterize oesophageal speech using words that are used during the rehabilitation of patients following total laryngectomy. The studied patients were under the care of the Bydgoszcz Laryngectomy Association (Bydgoszcz, Poland), where they underwent speech rehabilitation. These people spoke Polish, so all the analyzed words were spoken in Polish. The choice of words was determined by the rehabilitation process and depends on the specificity of the Polish language (Figure 1) [11]. This study aimed to analyze oesophageal speech using the division of the examined words into phonetic segments (phonetic syllables). Phonetic segments are changes in loudness between consecutive sounds in a stream of speech sounds. The center of the phonetic segment is the voice segment that differs in loudness level from the immediate surroundings. Its loudness is almost always greater than the loudness of the sound immediately before or after it. Each tested word was divided into phonetic segments, which were then parameterized with descriptors of time domain, frequency domain, cepstrum and mel coefficients. Separate parameterization of oesophageal speech, including the study of phonetic segments, is used to determine which of them show the most significant differences in descriptor values in relation to physiological speech (speech of healthy people). It means that for intelligent systems, we do not need to provide the whole word (as an input argument) but only its fragment in the form of a key phonetic segment of the studied word. This approach will lead to the acceleration of the functioning of speech recognition systems.

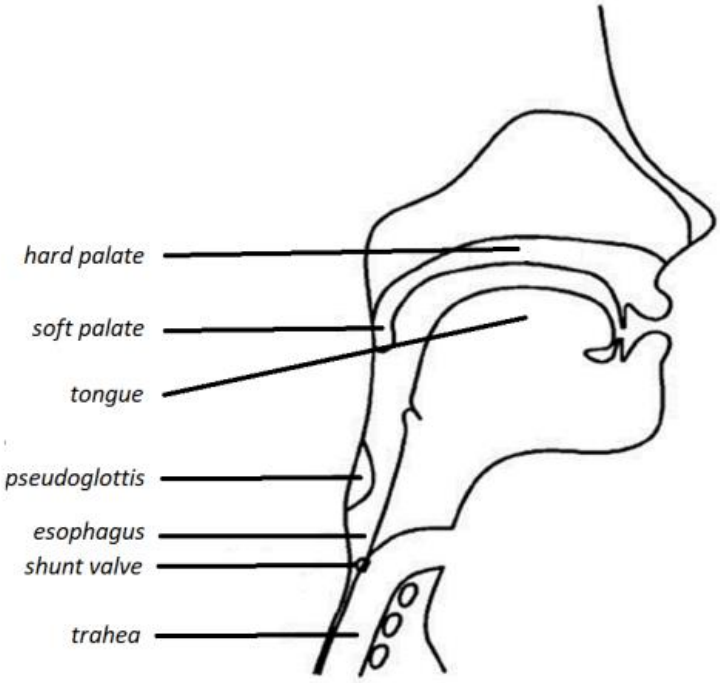

Figure 1. View following total laryngectomy. The pseudo-glottis is visible (own version).

The paper is divided into the following sections:

- Introduction: This section covers general issues related to laryngeal cancer, laryngectomy and available speech rehabilitation options, including their advantages and disadvantages;
- Materials and methods: This section lists the studied words, indicates the phonetic segments, and describes the recording conditions of the test words;
- Approach for obtaining feature vectors: This section discusses the time domain and spectrum domain descriptor definitions that were used during the research;
- Cepstrum analysis: This section discusses the definition of cepstrum and its interpretations;
- Mel-frequency cepstral coefficient (MFCC): This section of the paper discusses the MFCCs coefficients and how to extract them from a speech signal;
- Results: The results of the research are discussed here, and the effectiveness of the applied classification algorithms and learning methods in relation to the defined vectors of oesophageal speech features is indicated;
- Discussion: This is a place for summarizing the research and planning further research related to speech analysis.

2. Material and Methods

2.1. Material

A group of total laryngectomy patients participated in the study: three men aged 30–70 years and three women aged 30–60 years. The patients were undergoing speech rehabilitation related to learning oesophageal speech. Speech samples from healthy subjects were also studied from: three men aged 25–60 years and three women aged 20–50 years.

Speaker recognition was based on the classification algorithms used and the specific order of the feature vector. This is in line with [1,3,5].

The study was approved by the bioethics committee (KB 178/2020).

2.2. Methods

During the research, speech samples from the laryngectomy patients and healthy people were analyzed. People who had had a laryngectomy were undergoing speech rehabilitation aimed at learning oesophageal speech. For this reason, adequate words spoken, which are included in the rehabilitation programme, were examined. All words were spoken in Polish. In addition, speech samples were also taken from healthy people—the same spoken words were tested. The features of physiological speech provided a model (reference point) for comparing the values of speech descriptors in people after a laryngectomy. Taking into account the linguistic characteristics of the Polish language and the process of speech rehabilitation, the research covered the following words (spoken in Polish) [6]: a barrel, a bread roll, an egg, a package.

For this study, a set of scripts (a program) was written in the Octave environment (about 30 functions). The Octave environment is an alternative to Matlab but has the same capabilities and similar libraries. The use of this environment for speech analysis is one of the generally accepted research methods described in [12–14]. The functions created calculate indicator values or control the running of a specific function—e.g., changing a domain, loading a*.wav file for analysis, windowing signals, etc. The returned results are saved to various files, and from these files are created for the WEKA target (Figure 2). Of course, they take into account the configuration of the specific feature vector (treating the content of this feature vector as a set of descriptors).

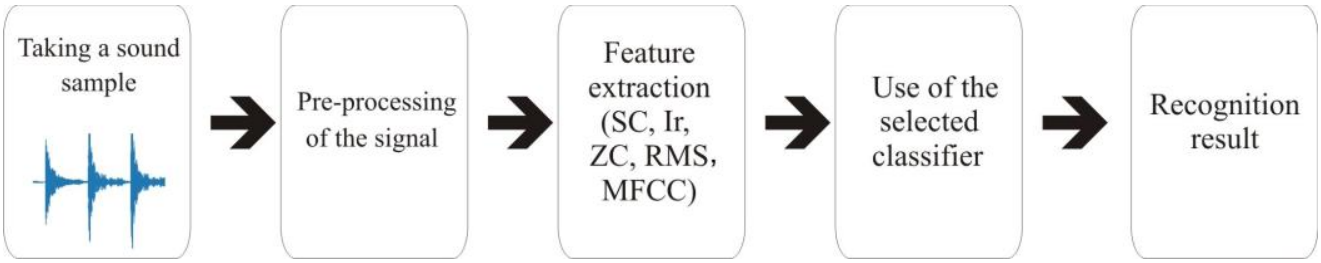

Figure 2. Diagram of experiments.

Each of the tested words has been divided into phonetic segments. In the case of the words "beczka" (eng. a barrel) and "paczka" (eng. a package), three phonetic segments were separated out. The words "bułka" (eng. a bread roll) and "jajko" (eng. an egg) were divided into two phonetic segments. The number of phonetic segments contained in these words is directly related to the features of the Polish language. Upon completion, four words from each person were used for the research. When segmenting the words (division into phonetic segments), ten samples (segments) were tested for each person. Each of them was tested independently. For speech signal analysis, the audio-physical characteristics (descriptors) must be extracted. The above segmentation of the words resulted from the features of the Polish language. Each phonetic segment of the tested word was independently parameterized by speech signal descriptors. The numerical values of the features obtained from each segment of the speech of the laryngectomized persons people were compared with the equivalents of words obtained from the healthy people. Table 1 presents the phonetic segments of the studied words. IPA notations for studied words:

- paczka—/paʧ̂ka/;
- jajko—/jæjkɔ/;
- beczka—/bɛʦ̂ʂka/;
- bułka—/buwka/.

Table 1. The list of the phonetic segments under study.

Researched Phonetic Segments				
In English	**In Polish**	**Seg 1**	**Seg 2**	**Seg 3**
a barrel	beczka	be	cz	ka
a bread roll	bułka	buł	ka	-
an egg	jajko	jaj	ko	-
a package	paczka	pa	cz	ka

Examples of the phonetic segments used are shown in Figures 3–6 using their time domains.

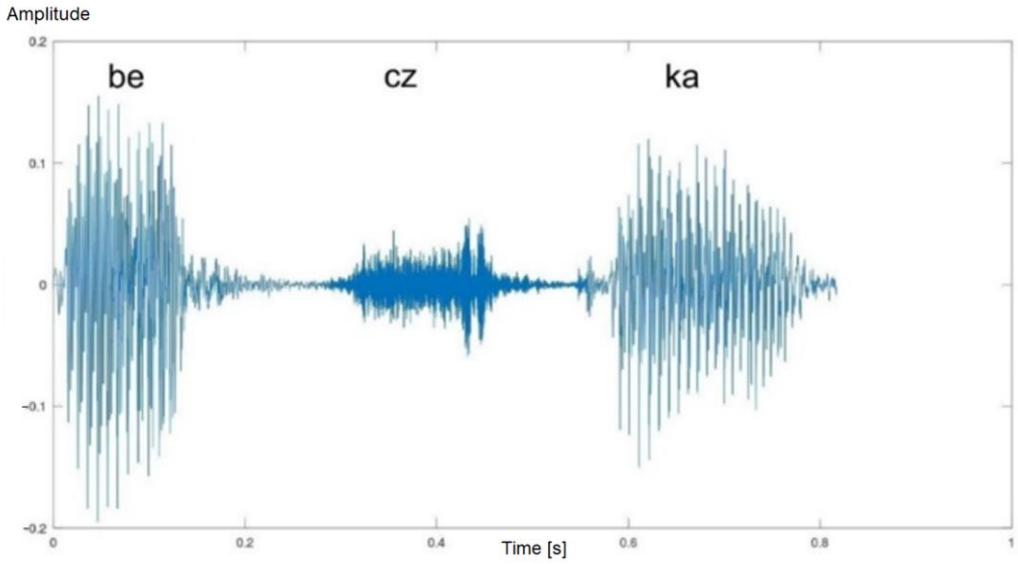

Figure 3. Time waveform of the word "beczka" (eng. a barrel)—healthy man. Three phonetic segments are visible.

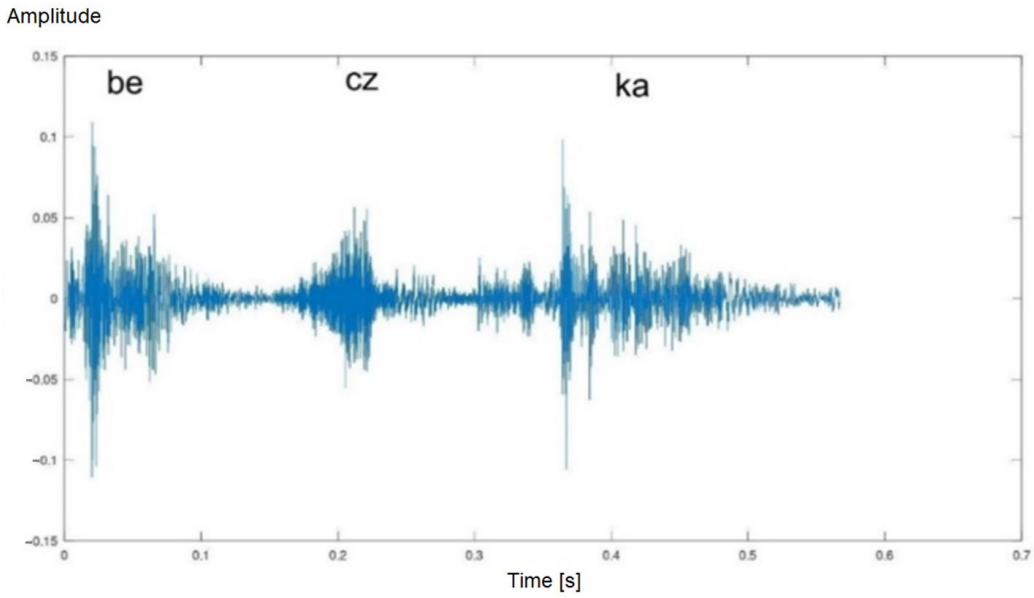

Figure 4. Time waveform of the word "beczka" (eng. a barrel)—laryngectomised man. Three phonetic segments are visible.

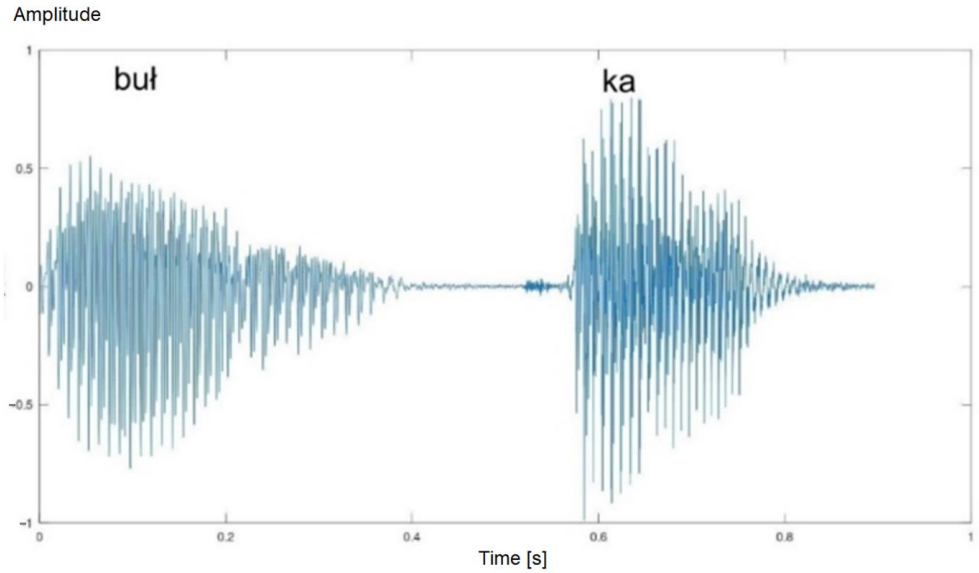

Figure 5. Time waveform of the word "bułka" (eng. a bread roll)—healthy man. Two phonetic segments are visible.

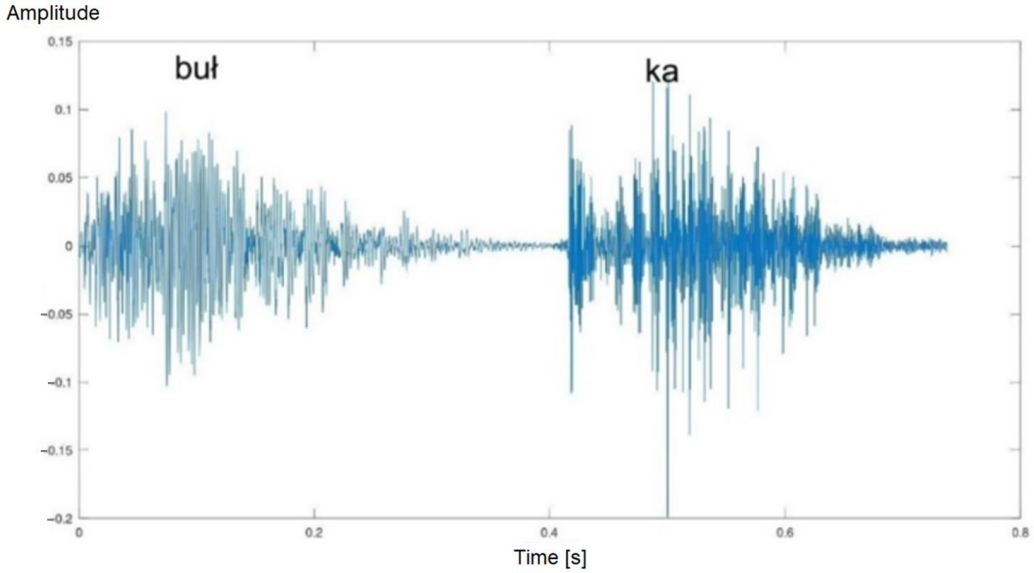

Figure 6. Time waveform of the word "bułka" (eng. a bread roll)—laryngectomised man. Two phonetic segments are visible.

The features of the physiological speech samples provided a baseline against which the features of oesophageal speech were compared. The recording of the speech samples took place at the Bydgoszcz Laryngectomy Association (city of Bydgoszcz, Poland). The recordings took place in a specially prepared room. An OMNITRONIC IM-1000 PRO condenser microphone was used for the recordings. All speech samples were recorded in WAV format with a sampling rate of 44,100 Hz and 16 bits/sample [11,15]. For the parameterization of the mentioned phonetic segments, widely used time- and frequency-domain descriptors were used. The Octave programming environment as well as the Praat and WaveSurfer computer programs were used to conduct the research. The WEKA package, with the classifiers implemented in it, was used to carry out the classification process.

Feature Vectors Obtained Approach

Speech-signal analysis can be defined as the process of extracting physical features from a speech signal (i.e., from sound samples). This process relies on the time- and frequency-domain parameterization mechanism to create a feature vector that allows the highest possible degree of object recognition. A feature vector defined in this way is a recognition pattern dedicated to a specific case of speech analysis, e.g., oesophageal speech. Both the time- and spectrum-domain descriptors were used for the research purposes [15,16]. Numerical values for the individual descriptors were obtained from each phonetic segment of the tested word.

2.3. Time Domain Descriptors

(1) *ZCR* (zero-crossing rate) is a measurement used to determine the ratio of zero crossings (the crossing of the OX axis). This is determined as the percentage of audio samples in a given fragment that change sign. The *ZCR* is defined by the following equation [17]:

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sgn[x_i(n)] - sgn[x_i(n-1)]| \quad (1)$$

where $sgn(*)$ is the function, i.e.,

$$sgn[x_i(n)] = \begin{cases} 1, x_i(n) \geq 0, \\ -1, x_i < 0. \end{cases} \quad (2)$$

In the research, the value of the *ZCR* descriptor was calculated in each phonetic segments of the speech of the healthy and laryngectomised people.

(2) Short-time energy (*STE*) is an audio descriptor from the MPEG-7 standard, also used in speech classification [6,18,19]. It describes the envelope of the signal. *STE* is the sum of squares computed in the time domain over the length of the test frame of the signal. The *STE* is expressed by the formula:

$$STE = \sum_{n=1}^{N} x^2(n) \quad (3)$$

where: $x(n)$—is the value of n-th sample, n—Index of the sample, N—signal length (total number of samples in the processing window, corresponding to the one phonetic segments).

(3) The signal mean value (*SMV*) descriptor expresses the average value of the input speech signal. Its value is estimated in the tested frame of the audio signal. It is calculated by summing the values of all samples and dividing by N. The *SMV* is given by:

$$SMV = \frac{1}{N} \sum_{n=1}^{N} x(n) \quad (4)$$

(4) Root mean square is the *RMS* value of a (periodic) signal, also known as nominal or continuous. This feature is widely used in speech parameterization. It is expressed by the formula [6,20]:

$$RMS = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x^2(n)} \quad (5)$$

where: N—signal length (total number of samples in the processing window, corresponding to the one phonetic segments), x—is the value of nth sample.

(5) Local minimum and maximum: The local maximum is the point at which the function changes from ascending to descending. Also, the local minimum is the point at which the function changes from descending to ascending. In the research, each phonetic

segment in the time domain was divided into 20 ms long windows. In each of these windows, the local minimum and maximum values were found.

2.4. Frequency Domain Descriptors

Spectral descriptors allow for the description of a speech signal with very high precision, which has a direct impact on the recognition of the speaker. These parameters are often used in speech recognition and improve the general recognition of the speaker. A very rich database of definitions of mathematical spectral parameter definitions can be found in [21–25]. Some of them have been implemented for the parameterization of phonetic segments of oesophageal speech following laryngectomy. The appropriate selection of spectral features enriches the definition of the oesophageal-speech pattern. The study of oesophageal speech in the frequency domain was carried out using signal windowing—which is the generally accepted standard when analyzing an audio signal. The length of the window for each phonetic segment was assumed to be 20 ms. Spectrum leakage was reduced by using a Hamming window and a 10 ms length overlap [2,18,19,26]. This approach to the study allowed us to obtain a spectrum matrix for the analyzed phonetic segment. Then, for each signal spectrum, the descriptors values were found and their average value was calculated. An example spectrum matrix for the first phonetic segment of the word "beczka" (Seg. 1 "be") spoken by a laryngectomised woman is presented in Figure 6.

This area of the cepstrum is the most different for oesophageal and physiological speech. This is due to the lack of the patient's larynx, which was removed by laryngectomy surgery. This area of the cepstrum in laryngectomees has flat characteristics—no larynx, no laryngeal tone.

For the research, the spectral-domain descriptors of oesophageal speech below were used:

(1) Spectral centroid (*SC*) is a way of describing the shape of the power spectrum. It shows whether the spectrum is dominated by low or high frequencies. This descriptor also refers to the timbre of the sound and allows the separation of tonal sounds from noise.

$$SC = \frac{\sum_{i=0}^{n} A(i) * i}{\sum_{i=0}^{n} A(i)} \quad (6)$$

where: $A(i)$ is amplitude of the i-th component (harmonic), i—index of the i-th partial.

(2) Irregularity of spectrum (*Ir*)

$$Ir = log\left(20 \sum_{i=2}^{N-1} \left| log \frac{A(i)}{\sqrt[3]{A(i-1) * A(i) * A(i+1)}} \right|\right) \quad (7)$$

2.4.1. Cepstrum Analysis

The cepstrum was obtained via the inverse Fourier transform applied to the logarithm of the signal spectrum. The domain of the cepstrum consists of pseudo-time values, which are called "quefrency". Low quefrency values represent slowly changing components of the logarithm of the spectrum logarithm, while high values represent to fast- changing components.

The cepstrum is obtained by the formula:

$$C(t)=IFFT[log\,|\,FFT[x(t)]\,|\,] \quad (8)$$

where: $x(t)$—analyzed windowed frame.

To obtain the cepstrum of speech, firstthe windowed signal frame needs to be transferred to the frequency domain using the fast Fourier transform (*FFT*) transform and then transferred back to the quefrency domain (an anagram of the word "frequency")—presented in the domain of the sample number [smpl] or time [s]. The resulting signal is known as a real cepstrum. Quefrency measured in seconds means that it does not indicate time but periods of frequency—peaks appearing in the cepstrum reveal periods of frequency that have harmonics in the spectrum. The quefrency domain is also called the pseudo-time domain. In the cepstrum, the low quefrencies contain information about the slowly changing features of the log-spectrum. The pictures below show the cepstrum of a laryngectomised woman and a healthy woman—the third phonetic segment "ka" of the word barrel (in Polish "beczka" in Polish, "beczka").

The Figures 7 and 8 illustrate the key differences for the representation of the cepstrum of physiological and oesophageal speech. It can be seen that the most significant differences are located between 50 and 100 smpl of the cepstrum. This area is characteristic of laryngeal speech. In laryngectomy patients, this range has a flat representation, as opposed to physiological speech. For this reason, the conducted research focused only on this range of the cepstrum.

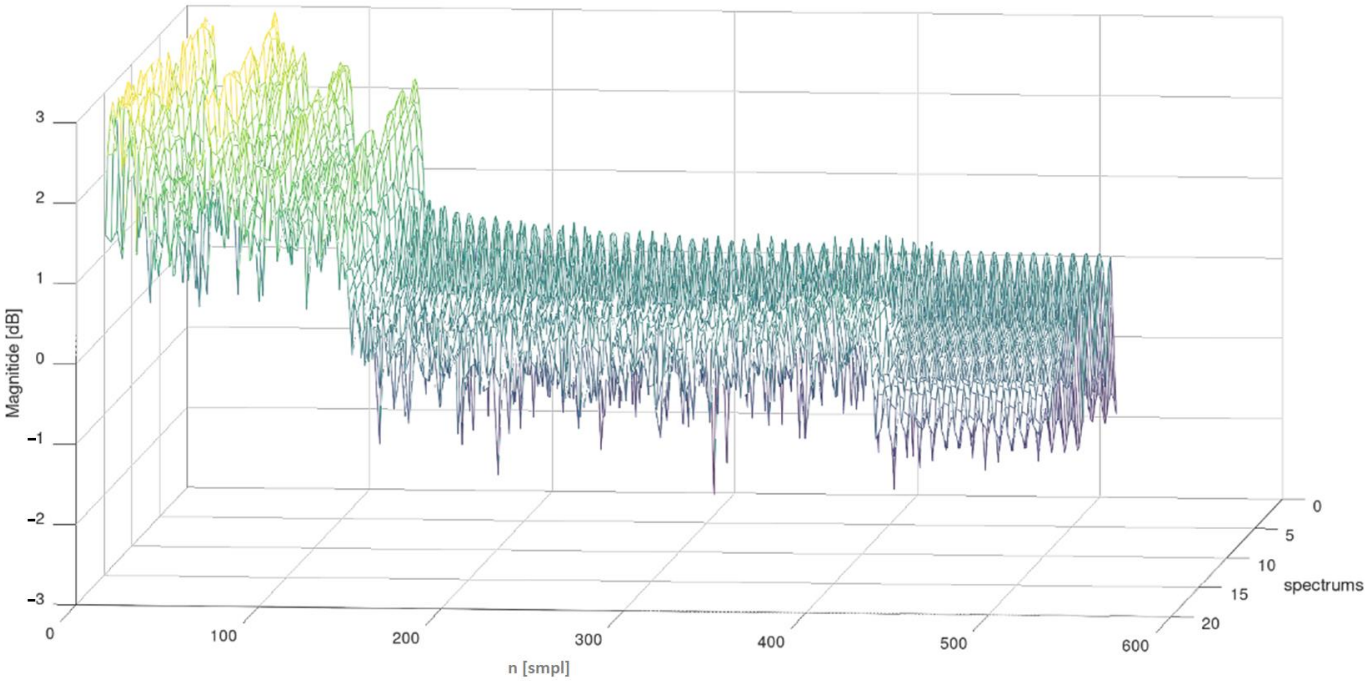

Figure 7. The matrix of spectra of the first phonetic segment of the word barrel (in Polish "beczka", Seg 1 "be")—spoken by a laryngectomised women.

2.4.2. Mel-Frequency Cepstral Coefficient

The *MFCC* feature vectors (mel frequency cepstral coefficients) are attributes describing the content of the cepstrum of the analyzed sound. It is a widely used method in speech recognition. The important point is that the *MFCC* analysis takes into account the perception of human hearing [24,25]. The *MFCC* parameter group is derived from the cepstrum of the signal represented in the mel scale. *MFCCs* encode spectra shape. The mel scale is characterized by the fact that it describes the perceptual distance between tones of different frequencies [26,27]. The relationship between the mel scale and the frequency scale is expressed as:

$$m = 2595 log_{10}\left(1 + \frac{f}{700}\right) \quad (9)$$

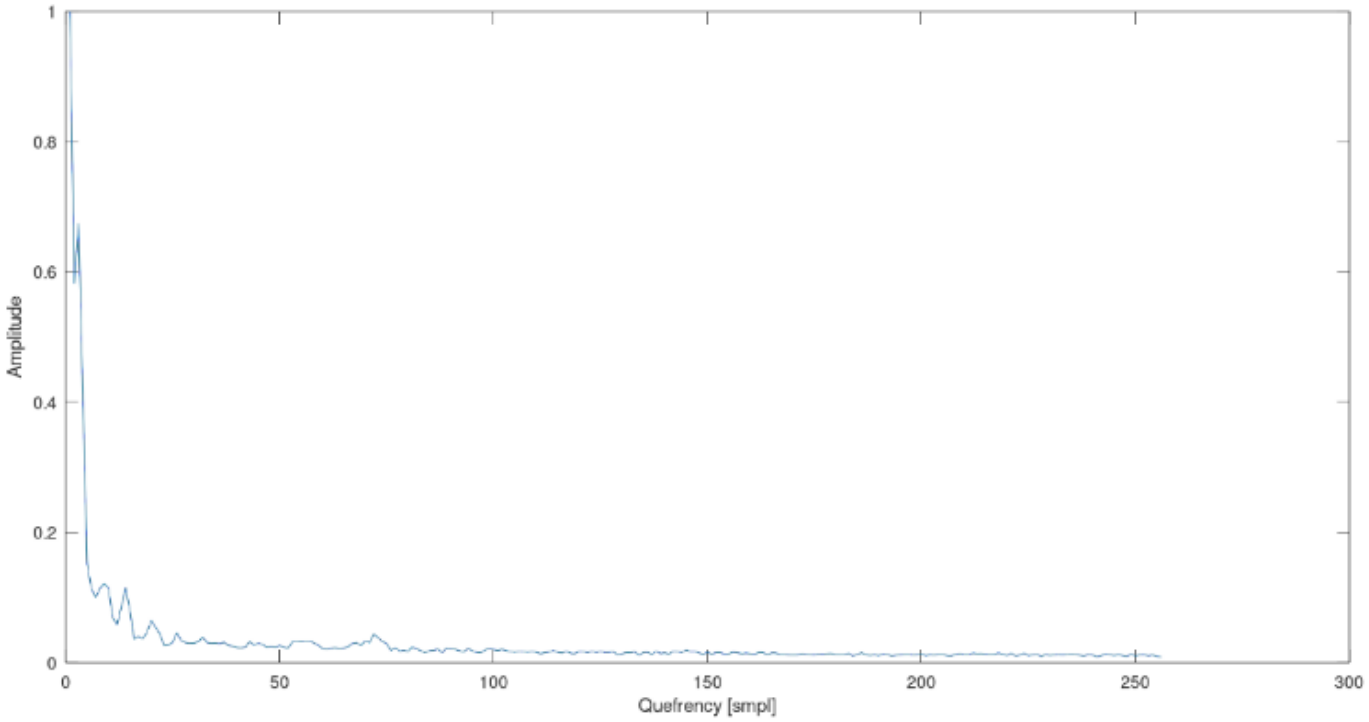

Figure 8. Cepstrum of the phonetic segment "ka", word barrel (in Polish "beczka")—woman who has undergone a laryngectomy.

Due to further operations on the mel-scale sound, a mel filter bank was used. The mel filter bank consists of bandpass filters with triangular amplitude characteristics overlapping with center frequencies 100 mels apart. Speech analysis typically uses 12 to 20 filters. The *MFCC* coefficients were used to analyse the phonetic segments of the speech of people after a laryngectomy. A bank of mel filters was used = 20. Figures 9 and 10 illustrate the *MFCC* feature set of the phonetic segment No. 1 ("buł") of the word bread roll (in Polish in Polish, "bułka"), spoken by a man who had undergone laryngectomy surgery.

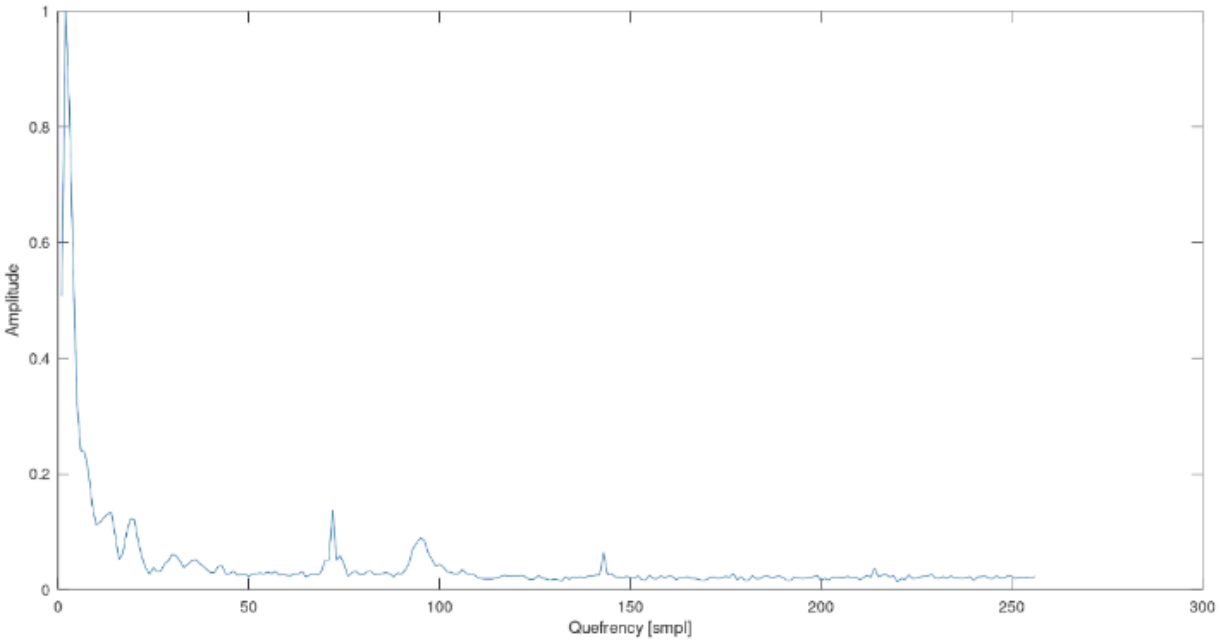

Figure 9. Cepstrum of the phonetic segment "ka", word barrel (in Polish "beczka")—the physiological speech of a healthy woman.

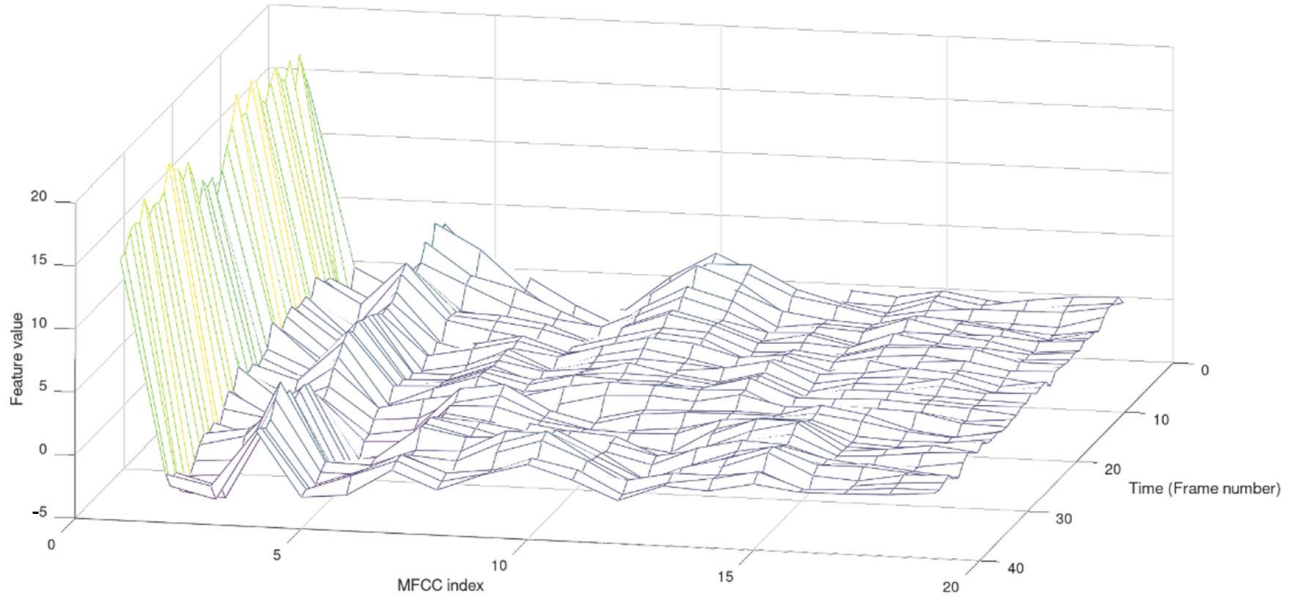

Figure 10. *MFCC* features of the Seg 1 ("buł") of word bread roll ("bułka")—man after laryngectomy.

3. Results

The study included people who had had total laryngectomy surgery. The purpose of the research was to define feature vectors extracted from the time and frequency domains. Two analytical approaches and a number of descriptors were used to parameterize the analyzed phonetic segments. The following speech-signal descriptors used:

- Zero-crossing rate (*ZCR*);
- Short-time energy *(STE*);
- Signal mean value (*SMV*);
- The root mean square (*RMS*);
- Local minimum and maximum;
- Spectral centroid (*SC*);
- Irregularity of spectrum (Ir);
- Cepstrum;
- *MFCC*—mel-frequency cepstral coefficient.

In the first analytical approach, the values of the *ZCR*, *STE*, *SMV* and *RMS* features were calculated over the length of the entire phonetic segment. The distribution of the values of the features was estimated on the basis of their minimum value, maximum value, mean value, and standard deviation. Below, the distribution values of the discussed features from the phonetic segments of the word "„barrel"" (in Polish "beczka" in Polish, "beczka"), have been presented (Tables 2–4).

Table 2. The distribution of the values of the features: the first phonetic segments „be"—the word "bec-zka".

	ZCR	STE	SMV	RMS
Laryngectomized persons				
average	431.67	251.13	0.034	0.181
min	169.00	113.40	0.016	0.126
max	621.00	376.75	0.055	0.234
SD	162.65	110.80	0.015	0.042

Table 2. *Cont.*

	ZCR	STE	SMV	RMS
		Healthy persons		
average	209.75	606.79	0.083	0.286
min	80.00	374.83	0.064	0.252
max	269.00	854.15	0.107	0.327
SD	87.37	209.28	0.018	0.031

SD—standard deviation.

Table 3. The distribution of the values of the features: the second phonetic segments „cz"—the word "beczka".

	ZCR	STE	SMV	RMS
		Laryngectomized persons		
average	543.67	152.25	0.039	0.187
min	214.00	37.93	0.012	0.110
max	916.00	398.20	0.078	0.280
SD	274.73	128.06	0.028	0.071
		Healthy persons		
average	464.50	384.22	0.128	0.322
min	148.00	108.09	0.045	0.211
max	1311.00	908.49	0.350	0.591
SD	564.95	376.96	0.148	0.181

SD—standard deviation.

Table 4. The distribution of the values of the features: the third phonetic segments „ka"—the word "beczka".

	ZCR	STE	SMV	RMS
		Laryngectomized persons		
average	789.33	284.26	0.026	0.156
min	383.00	121.42	0.010	0.102
max	1217.00	625.81	0.043	0.207
SD	303.23	186.58	0.013	0.043
		Healthy persons		
average	339.25	766.80	0.073	0.268
min	306.00	571.78	0.053	0.231
max	378.00	983.39	0.095	0.309
SD	38.13	225.05	0.019	0.036

SD—standard deviation.

We decided to use the k-NN classifier and the cross-validation method to classify the data. Cross-validation meant that the dataset was divided into K subsets. Then, in order, each of these subsets was treated as a test set and the others as training sets. This analysis was performed k times. The k results obtained were averaged to obtain a single result. The choice of the parameter k depends on the size of the datasets and their type. For large datasets, k = 3 is usually used to reduce the number of model adaptations. For smaller datasets, larger values of k are usually used in order not to deplete the training set too much, which could result in low model quality. In this case, k = 10 is most often used. Due to the small population of people who had had laryngectomy surgery, we decided to use k = 10. Both the chosen classifier and the cross-validation method have already been used by the authors in previous studies related to audio analysis. Recognition results for all phonetic segments of the laryngectomised people and the healthy people using the *ZCR*, *STE*, *SMV* and *RMS* descriptors are summarized in the tables below (Tables 5–7).

Table 5. Error matrix for the classification of the first phonetic segment. Used k-NN, cross-validation method (k = 10). General recognition 93.75%.

a	b	Classified as
93.75	6.25	a = laryngectomized
6.25	**93.75**	b = healthy

Table 6. Error matrix for the classification of the second phonetic segment. Used k-NN, cross-validation method (k = 10). General recognition 62.5%.

a	b	Classified as
56.25	43.75	a = laryngectomized
31.25	**68.75**	b = healthy

Table 7. Error matrix for the classification of the third phonetic segment. Used k-NN, cross-validation method (k = 10). General recognition 87.5%.

a	b	Classified as
100	0	a = laryngectomized
25	**75**	b = healthy

Error matrices (also called confusion matrix) should be interpreted as:

1. Markings "a" and "b" as a group of examined people (laryngectomized and healthy);
2. Shaded cells in the table contain the correct classification. For example, in Table 5, in1st row, 93.75% means the correct classification of samples from laryngectomised people, and 6.25% means incorrect classification.

It should be noted that in this case, the laryngectomized people were recognized with 100% accuracy—with 75% accuracy in the recognize of healthy people. Due to the adopted procedure for analyzing the frequency domain of phonetic segments (described in point Section 2.4 "Frequency Domain Descriptors"), the features of Ir and Br (computed for each spectrum of this matrix) were obtained from the spectral matrix for a single signal and then their average value was calculated. This means that Ir and Br for a specific phonetic segment is the average of the values of these features obtained from the obtained spectrums (an example spectrum matrix is shown in Figure 6). The same approach (calculating the value of the average feature from the spectral matrix) was also used in the analysis of the *MFCC* coefficients. Below, the distribution values of the discussed features for the phonetic segments of the word "package" have been shown (in Polish "in Polish", "paczka") (Tables 8–10).

Table 8. The first phonetic segments „pa"—descriptor values—the word "paczka".

	Br	Ir
Laryngectomized persons		
average	115.645	3.79
min	96.463	3.67
max	129.080	3.83
SD	13.414	0.07
Healthy persons		
average	124.615	3.84
min	118.576	3.84
max	132.847	3.85
SD	6.279	0.01

SD—standard deviation.

Table 9. The second phonetic segments „cz"—descriptor values—the word "paczka".

	Br	Ir
Laryngectomized persons		
average	105.538	3.69
min	78.601	3.58
max	114.724	3.80
SD	15.127	0.10
Healthy persons		
average	114.653	3.75
min	106.213	3.69
max	128.502	3.80
SD	12.089	0.06

SD—standard deviation.

Table 10. The third phonetic segments „ka"—descriptor values—the word "paczka".

	Br	Ir
Laryngectomized persons		
average	112.193	3.74
min	99.956	3.68
max	120.649	3.83
SD	7.982	0.06
Healthy persons		
average	146.588	3.81
min	123.233	3.78
max	159.358	3.86
SD	15.969	0.03

SD—standard deviation.

Recognition results for all phonetic segments in laryngectomised people and the healthy people using the Ir and Br descriptors are presented in the Tables 11–13.

Table 11. Error matrix for the classification of the first phonetic segment. Used k-NN, cross-validation method (k = 10). General recognition 50%.

a	b	Classified as
43.75	56.25	a = laryngectomized
43.75	**56.25**	b = healthy

Table 12. Error matrix for the classification of the second phonetic segment. Used k-NN, cross-validation method (k = 10). General recognition 43.75%.

a	b	Classified as
43.75	56.25	a = laryngectomized
56.25	**43.75**	b = healthy

Table 13. Error matrix for the classification of the third phonetic segment. Used k-NN, cross-validation method (k = 10). General recognition 87.5%.

a	b	Classified as
100	0	a = laryngectomized
25	**75**	b = healthy

Having analyzed the results, it can be concluded that the two-element vector of spectral features (Ir and Br) did not provide satisfactory classification effects. This is especially true of the second phonetic segment, where the correct diagnosis recognition was below 50% of the general classification. The satisfactory result was obtained only for the third phonetic segment, which was identical to the vector of time-domain features (*ZCR, STE, SMV, RMS*). In the second analytical approach, the values of the *ZCR*, *STE*, *SMV*, *RMS*, Ir and Br features were calculated in each window (windowing is 20 ms) of the phonetic segment. The distribution of each feature in each phonetic segment was observed. In this case, the mean value of the descriptor from all spectra was not calculated. The examples of the *RMS* features and Br features distribution for the first phonetic segment of the word barrel (in Polish "beczka" in Polish, "beczka") are shown on Figures 11 and 12.

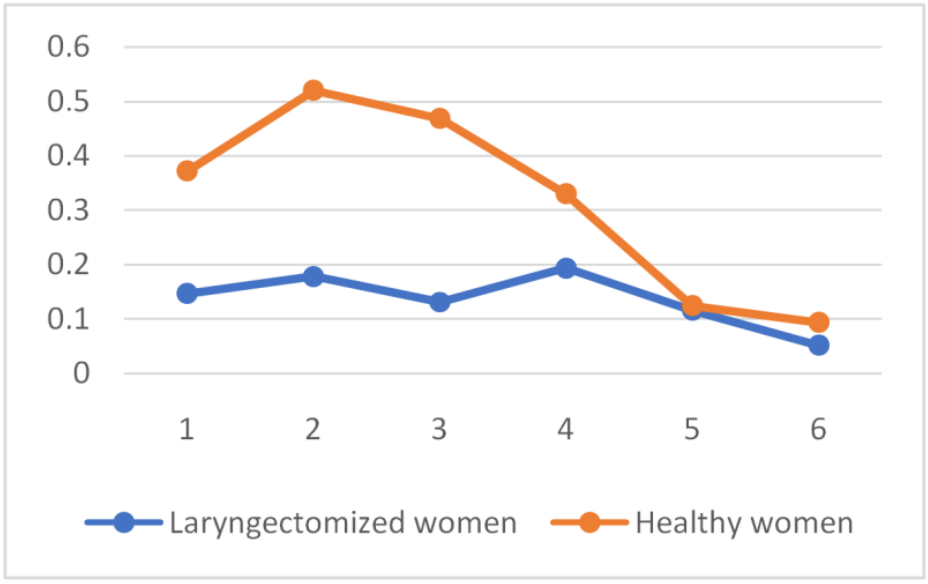

Figure 11. RMS feature distribution for the first phonetic segment of the word „barrel" (in Polish „beczka").

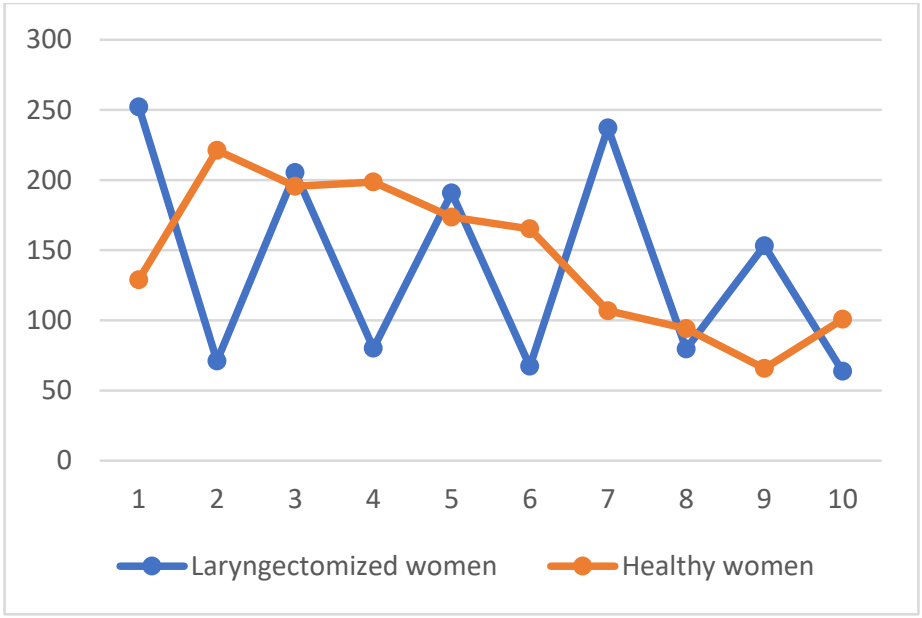

Figure 12. Br features distribution for the first phonetic segment of the word „barrel" (in Polish „beczka").

Taking into account the different length of individual phonetic segments, the number of feature component vectors (i.e., the first n signal windows) was standardized:

1. For *ZCR*, *STE*, *SMV*, *RMS* descriptors:
   - Seg. 1: 6 features;
   - Seg. 2: 3 features;
   - Seg. 3: 6 features;
2. For Ir and Br descriptors:
   - Seg. 1: 10 features;
   - Seg. 2: 4 features;
   - Seg. 3: 12 features.

Moreover, a comparison of the local minimum and maximum distributions in each window of the waveform of phonetic segments was made. The example of the local minimum and local maximum distributions for the first phonetic segment of the word "barrel" (in Polish "beczka" in Polish, "beczka") is shown on the fin Figure 13.

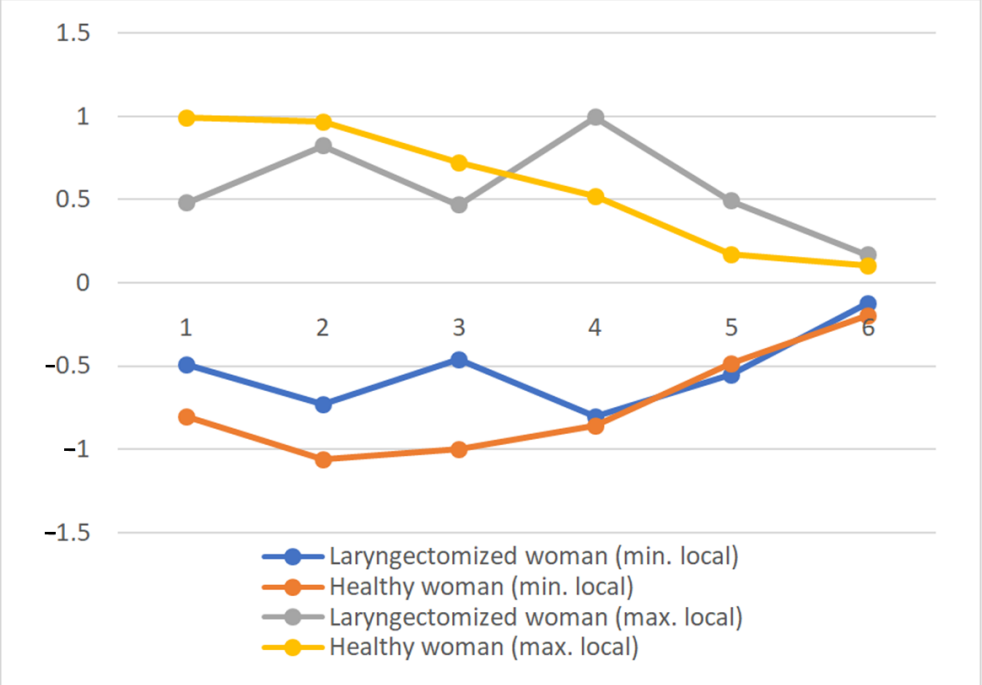

Figure 13. Local minimum and local maximum distributions for the first phonetic segment of the word „barrel" (in Polish "beczka").

The table below (Table 14) presents the remaining results of the research carried out for each phonetic segment with the use of the described feature vectors. All the classification results presented in the table apply to k-NN and the cross-validation method for k = 10.

Table 14. Summary of the classification results.

Segment	Features Vector	General Recognition		Laryngectomized		Healthy	
		Correctly Classified	Incorrectly Classified	Correctly Classified	Incorrectly Classified	Correctly Classified	Incorrectly Classified
Seg 1	MFCC (19 coefficients)	79.42%	20.58 %	94.12%	5.88%	64.7%	35.29%
Seg 2	MFCC (19 coefficients)	87.5%	12.5%	87.5%	12.5%	87.5%	12.5%

Table 14. *Cont.*

Segment	Features Vector	General Recognition		Laryngectomized		Healthy	
		Correctly Classified	Incorrectly Classified	Correctly Classified	Incorrectly Classified	Correctly Classified	Incorrectly Classified
Seg 3	MFCC (19 coefficients)	81.25%	18.75%	87.5%	12.5%	75%	25%
Seg 1	Cepstrum (40 features)	84.4%	15.6%	93.75%	6.25%	75%	25%
Seg 2	Cepstrum (40 features)	40.63%	59.37%	56.25%	43.75%	25%	75%
Seg 3	Cepstrum (40 features)	62.5%	37.5%	62.5%	37.5%	62.5%	37.5%
Seg 1	local maximum distributions (6 features)	85.7%	14.3%	71.43%	28.57%	100%	0
Seg 2	local maximum distributions (3 features)	64%	36%	61.54%	38.46%	66.66%	33.33%
Seg 3	local maximum distributions (6 features)	81.25%	18.75%	62.5%	37.5%	100%	0
Seg 1	local minimum distributions (6 features)	57.15%	42.85%	50%	50%	64.3%	35.7%
Seg 2	local minimum distributions (3 features)	75%	25%	83.33%	16.66%	66.66%	33.33%
Seg 3	local minimum distributions (6 features)	81.25%	18.75%	62.5%	37.5%	100%	0
Seg 1	ZCR features distributions (6 features)	87.5%	12.5%	81.25%	18.75%	93.75	6.25%
Seg 2	ZCR features distributions (3 features)	60%	40%	66.66%	33.33%	53.33%	46.66%
Seg 3	ZCR features distributions (6 features)	81.25%	18.75%	87.5%	12.5%	75%	25%
Seg 1	STE features distributions (6 features)	68.75%	31.25%	62.5%	37.5%	75%	25%
Seg 2	STE features distributions (3 features)	70.83%	29.16%	83.33%	16.66%	58.33%	41.66%
Seg 3	STE features distributions (6 features)	81.25%	18.75%	87.5%	12.5%	75%	25%
Seg 1	SMV features distributions (6 features)	75%	25%	75%	25%	75%	25%
Seg 2	SMV features distributions (3 features)	63.33%	36.66%	66.66%	33.33%	60%	40%
Seg 3	SMV features distributions (6 features)	81.25%	18.25%	87.5%	12.5%	75%	25%
Seg 1	RMS features distributions (6 features)	78.13%	21.87%	68.75%	31.25%	87.5%	12.5%
Seg 2	RMS features distributions (3 features)	63.33%	36.66%	66.66%	33.33%	60%	40%
Seg 3	RMS features distributions (6 features)	81.25%	18.75%	87.5%	12.5%	75%	25%
Seg 1	Br features distribution (10 features)	60%	40%	60%	40%	60%	40%
Seg 2	Br features distribution (4 features)	63.33%	36.66%	66.66%	33.33%	60%	40%
Seg 3	Br features distribution (12 features)	68.75%	31.25%	100%	0%	62.5%	37.5%
Seg 1	Ir features distribution (10 features)	62.5%	37.5%	37.5%	62.5%	87.5%	12.5%
Seg 2	Ir features distribution (4 features)	43.75%	56.25%	37.5%	62.5%	50%	50%
Seg 3	Ir features distribution (12 features)	62.5%	37.5%	50%	50%	75%	25%

4. Discussion

This paper proposes a method for analyzing oesophageal speech based on the analysis of phonetic segments determined by changes in loudness between successive sounds in the speech signal, the centre of which differs in loudness level from the nearest neighbor in the signal. As a result of the experiments, the most effective descriptors for parameterization of oesophageal speech were identified. These are the *MFCC* coefficients, the local minimum and maximum indications, and the *ZCR*. Furthermore, it was found that the most important features of oesophageal speech occur mainly in the first phonetic segment. The table above presents the results of the classification of the phonetic segments (Seg. 1, Seg. 2 and Seg. 3) of the oesophageal and physiological speech. They provide a set of patterns that can be used to recognize the features of oesophageal speech—especially implemented during speech rehabilitation following laryngectomy surgery. To carry out the study, the WEKA package was used, from which the k-NN classifier and the cross-validation method for k = 10 were selected. Using the aforementioned classifier, all proposed featurevectors of features were tested. Since k-NN was used in each case of classification, it is possible to evaluate the effectiveness of the proposed feature vectors. It should be noted that the presented classification is one of the possible solutions, because it is plausible to combine the components of the features contained in the above feature vectors. However, other classification algorithms and rules (e.g., random forest or decision tables) can be applied, yet these will certainly provide different results. It was also shown that the effectiveness of the proposed feature vectors was different: beginning from good effectiveness to very poor effectiveness, and consequently did not provide any practical applications. Nonetheless, this knowledge is not without significance, as it allows us to exclude the sense of using ineffective descriptors for the classification of oesophageal speech in rehabilitated people. Examples of very poor classification are the cepstrum descriptors used for the second phonetic segment (general recognition 40.,63%) and the distribution of four Ir descriptors used for the second phonetic segment (general recognition 43.75%). In general, it should be pointed out that good classification efficiency is provided by the *MFCC* descriptors and the distribution of local maximum and minimum (general recognition is above 80%). The *MFCC* descriptors also showed a good efficiency in the general recognition of oesophageal and physiological speech during whole- word analysis (without phonetic segmentation), which has been presented in detail [6]. Moreover, the distribution of the *ZCR* descriptors for the first phonetic segment provides an overall classification score of almost 90%, being a very good result. It is worth noting that the analysis of the distribution of local minima and maxima for individual recognition of physiological speech reached 100% in three cases.

The present study used a speech analysis approach (including syllable analysis of oesophageal speech) according to [28] and developed methods previously described by [29], which is in line with the main research trend. This allowed us to develop a group of tools based on coefficient analysis in the time domain, frequency domain, cepstrum, and *MFCC* to study the physical features of oesophageal speech in people after total laryngectomies, and test them on the results of real healthy people, as well as the future development of this group of technologies to support early diagnosis and even prevention, and to compare our results with those of other Rothera research groups [28–32].

In view of the planned use of the research results to improve early diagnosis and control of the treatment, rehabilitation, and care process in a group of patients after laryngeal surgery, our approach based on the analysis of phonetic segments of spoken oesophageal speech words may be helpful in improving the speech-rehabilitation process. The results of the study may be of relevance to [4,11]. Our solution is at various stages of research and the final therapeutic version of our solution is being developed in collaboration with specialists from a renowned centre: Bydgoszcz Laryngectome Association (Bydgoszcz, Poland) [11,33,34], which will allow us to better tailor the solution to the demands of patient group and its specificities.

The large number of related features indicates the need and opportunity for advanced analyses to extract the most important ones. Relevant findings include:

- Indication that the classification based on a temporal feature vector is more effective than a frequency feature vector;
- Identifying the first phonetic segment as the part of the word under study with the highest number of features is relevant for classification purposes—of all the phonetic segments, the first segment showed the highest classification performance—especially on the temporal feature index;
- High classification performance of the feature vector containing *MFCC* coefficients (across all three segments, the average recognition performance is about 83%);
- High classification performance of the feature vector for one phonetic segment: about 84% overall recognition performance;
- High classification efficiency resulting from the analysis of local minima and maxima: about 86% for the 1st phonetic segment and about 81% for the 3rd phonetic segment.

The results will allow for a future focus on the analysis of the aforementioned phonetic segments and the search for new feature vectors.

We have taken the definition of the term 'phonetic segment' from the approach in [35–38] as we develop tools to support this research, diagnostic and therapeutic method. This will allow for a better understanding of the processes involved in speech rehabilitation related to the teaching of replacement speech (oesophageal speech), where, in the absence of the larynx, the role of the sound source is played by the oesophageal-mandibular folds, i.e., where the source of the speech signal is the pseudo pharynx. The research conducted will prepare the ground for the further development of this method, i.e., the analysis of oesophageal speech, where the word under study is divided into phonetic segments before feature extraction.

Directions for Further Research

Further analysis of oesophageal speech with a particular focus on defining such a vector of features that will further support the process of speech rehabilitation is planned. Therefore, it is necessary to obtain samples of oesophageal speech from people who participate in each of the five levels of speech rehabilitation [10,35–37]. The control and appropriate selection of descriptors will enable the definition of feature vectors that will be appropriate for each stage of the implemented rehabilitation. Controlling the dynamics of changes in descriptor values will make it possible to improve the speech-rehabilitation process of people who have undergone laryngectomy [38,39]. The above assumptions are the aim of further research on oesophageal speech.

Possibilities of contact with people deprived of the function of physiological speech are also implemented using the "silent speech interface (SSI)". This implementation consists in the use of device-enabling voice communication without the use of sound. SSI systems are a type of electronic lip-reading device. SSIs were created using ultrasound and an optical camera to analyze tongue and lip movements. A very broad description and application of SSI systems is included in [40], and our experience is enhanced by [41–50].

5. Conclusions

In general, it should be stated that by adopting the applied research methodology more often, better classification results are obtained by analyzing Seg. 1 and Seg. 2, especially for time-domain descriptors. Also, good classification results have been obtained using the MFCC descriptors. The worst results of the research were recorded using the Br and Ir descriptors.

Author Contributions: Conceptualization, K.T., I.R. and D.M.; methodology, K.T.; software, K.T.; validation, K.T., I.R. and D.M.; formal analysis, K.T.; investigation, K.T.; resources, K.T.; data curation, K.T.; writing—original draft preparation, K.T., I.R. and D.M.; writing—review and editing, K.T., I.R. and D.M.; visualization, K.T.; supervision, K.T., I.R. and D.M.; funding acquisition, I.R. All authors have read and agreed to the published version of the manuscript.

Funding: The work presented in the paper has been financed under grant to maintain the research potential of Kazimierz Wielki University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Report: National Health Fund Headquarters, The Analysis and Innovation Department, Warszawa, Poland. July 2021. Available online: https://ezdrowie.gov.pl/pobierz/nfz-o-zdrowiu-choroby-odtytoniowe-popr (accessed on 28 January 2023).
2. Guidotti, L.; Negroni, D.; Sironi, L.; Stecco, A. Neural Correlates of Esophageal Speech: An fMRI Pilot Study. *J. Voice* **2022**, *36*, 288.e1–288.e14. [CrossRef] [PubMed]
3. Doyle, P.C.; Damrose, E.J. Has Esophageal Speech Returned as an Increasingly Viable Postlaryngectomy Voice and Speech Rehabilitation Option? *J. Speech Lang. Hear. Res.* **2022**, *65*, 4714–4723. [CrossRef] [PubMed]
4. Hong, S.-W.; Chan, R.W. Acoustic Analysis of Taiwanese Tones in Esophageal Speech and Pneumatic Artificial Laryngeal Speech. *J. Speech Lang. Hear. Res.* **2022**, *65*, 1215–1227. [CrossRef] [PubMed]
5. Kresic, S.; Veselinovic, M.; Mumovic, G.; Mitrović, S.M. Possible factors of success in teaching esophageal speech. *Med. Rev.* **2015**, *68*, 5–9. [CrossRef]
6. Sokal, W. Possibilities of Verbal Communication in Patients after Complete Removal of the Larynx. Ph.D. Dissertation, Poznan University of Medical Science, Poznań, Poland, 2011.
7. Tyburek, K. Parameterisation of human speech after total laryngectomy surgery. *Comput. Speech Lang.* **2022**, *72*, 101313. [CrossRef]
8. Ezzine, K.; Di Martino, J.; Frikha, M. Intelligibility Improvement of Esophageal Speech Using Sequence-to-Sequence Voice Conversion with Auditory Attention. *Appl. Sci.* **2022**, *12*, 7062. [CrossRef]
9. Uloza, V.; Maskeliunas, R.; Pribuisis, K.; Vaitkus, S.; Kulikajevas, A.; Damasevicius, R. An Artificial Intelligence-Based Algorithm for the Assessment of Substitution Voicing. *Appl. Sci.* **2022**, *12*, 9748. [CrossRef]
10. Zenga, J.; Goldsmith, T.; Bunting, G.; Deschler, D.G. State of the art: Rehabilitation of speech and swallowing after total laryngectomy. *Oral Oncol.* **2018**, *86*, 38–47. [CrossRef]
11. Sinkiewicz, A. Laryngeal Cancer. In *A Guide for Patients, Speech Therapists and Doctors*; Polish Society of Laryngectomes; Polish Society of Otolaryngologists; Head and Neck Surgeons: Poznań, Poland, 1999.
12. Amin, T.B.; Mahmood, I. Speech Recognitionusing Dynamic Time Warping. In Proceedings of the 2008 2nd International Conference on Advances in Space Technologies, Islamabad, Pakistan, 29–30 November 2008; pp. 74–79. [CrossRef]
13. Vyas, M. A Gaussian Mixture Model Based Speech Recognition System Using Matlab. *Signal Image Process. Int. J.* **2013**, *4*, 109. [CrossRef]
14. Patel, K.; Prasad, R.K. Speech Recognition and Verification Using MFCC & VQ. *Int. J. Emerg. Sci. Eng.* **2013**, *1*, 7.
15. Shim, H.J.; Jang, H.R.; Shin, H.B.; Ko, D.H. Cepstral, Spectral and Time-Based Analysis of Voices of Esophageal Speakers. *Folia PhoniatrLogop.* **2015**, *67*, 90–96. [CrossRef] [PubMed]
16. Lachhab, O.; Di Martino, J.; Elhaj, E.I.; Hammouch, A. A preliminary study on improving the recognition of esophageal speech using a hybrid system based on statistical voice conversion. *Springerplus* **2015**, *4*, 644. [CrossRef]
17. Giannakopoulos, T.; Pikrakis, A. *Introduction to Audio Analysis: A Matlab Approach*; Academic Press Books—Elsevier: Amsterdam, The Netherlands, 2014.
18. Tyburek, K.; Prokopowicz, P.; Kotlarz, P.; Repka, M. Comparison of the Efficiency of Time and Frequency Descriptors Based on Different Classification Conceptions. In Proceedings of the Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, 14–18 June 2015.
19. Tyburek, K.; Cudny, W.; Kosiński, W. Pizzicato sound analysis of selected instruments in the freguency domain. *Image Process. Commun.* **2006**, *11*, 53–57.
20. Titze, I.R.; Sundberg, J. Vocal intensity in speakers and singers. *J. Acoust. Soc. Amer.* **1992**, *91*, 2936–2946. [CrossRef] [PubMed]
21. Lindsay, A.T.; Burnett, I.; Quackenbush, S.; Jackson, M. *Fundamentals of Audio Descriptions in Introduction to Mpeg-7: Multimedia Content Description Interface*; Wiley and Sons Ltd.: Hoboken, NJ, USA, 2002; pp. 283–298.

22. Tyburek, K.; Kotlarz, P. An expert system for automatic classification of sound signals. *J. Telecommun. Inf. Technol.* **2020**, *2*, 86–90. [CrossRef]
23. Prokopowicz, P.; Mikołajewski, D.; Tyburek, K.; Mikołajewska, E. Computational gait analysis for post-stroke rehabilitation purposes using fuzzy numbers, fractal dimension and neural networks. *Bull. Pol. Acad. Sci. Tech. Sci.* **2020**, *68*, 191–198. [CrossRef]
24. Marechal, C.; Mikołajewski, D.; Tyburek, K.; Prokopowicz, P.; Bougueroua, L.; Ancourt, C.; Węgrzyn-Wolska, K. Survey on AI-Based Multimodal Methods for Emotion Detection. In *High-Performance Modelling and Simulation for Big Data Applications*; Kołodziej, J., González-Vélez, H., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11400.
25. Balemarthy, S.; Sajjanhar, A.; Zheng, J.X. Our Practice of Using Machine Learning to Recognize Species by Voice. *arXiv* **2018**, arXiv:1810.09078.
26. Fayek, H. Speech Processing for Machine Learning: Filter Banks, Mel-Frequency Cepstral Coefficients (mfccs) and What's in between. Available online: https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html (accessed on 28 January 2023).
27. Dobres, R.; Lee, L.; Stemple, J.C.; Kummer, A.W.; Kretschmer, L.W. Description of Laryngeal Pathologies in Children Evaluated by Otolaryngologists. *J. Speech Hear. Disord.* **1990**, *55*, 526–532. [CrossRef] [PubMed]
28. Liu, H.; Wana, H.; Wang, S.; Wang, X. Acoustic characteristics of Mandarines ophagealspeech. *J. Acoust. Soc. Am.* **2005**, *118*, 1016. [CrossRef]
29. Vojtech, J.M.; Chan, M.D.; Shiwani, B.; Roy, S.H.; Heaton, J.T.; Meltzner, G.S.; Contessa, P.; De Luca, G.; Patel, R.; Kline, J.C. Surface Electromyography-Based Recognition, Synthesis, and Perception of Prosodic Subvocal Speech. *J. Speech Lang. Hear. Res.* **2021**, *64*, 2134–2153. [CrossRef]
30. Wang, H.; Roussel, P.; Denby, B. Improving ultrasound-based multimodal speech recognition with predictive features from representation learning. *JASA Express Lett.* **2021**, *1*, 015205. [CrossRef]
31. Allegra, E.; La Mantia, I.; Bianco, M.R.; Drago, G.D.; Le Fosse, M.C.; Azzolina, A.; Grillo, C.; Saita, V. Verbal performance of Total laryngectomized patients rehabilitated with esophageal speech and tracheoesophageal speech: Impacts on patient quality of life. *Psychol. Res. Behav. Manag.* **2019**, *12*, 675–681. [CrossRef] [PubMed]
32. Wszołek, W.; Modrzejewski, M.; Przysiężny, M. Acoustic analysis of esophageal speech in patients after tallaryngectomy. *Arch. Acoust.* **2014**, *32*, 151–158.
33. Wamka, M.; Mackiewicz-Nartowicz, H.; Sinkiewicz, A. Nursing care of patients after laryngeal surgery. *Surg. Angiol. Nurs.* **2018**, *4*, 136–140.
34. Mackiewicz-Nartowicz, H.; Mackiewicz-Milewska, M. Epidemiology, Etiology and Diagnosis of Laryngeal Cancer. In *Patient after Larynx Surgery*; Sinkiewicz, A., Ed.; Bydgoszcz Laryngectomy Association: Bydgoszcz, Poland, 2009.
35. Botinis, A.; Granström, B.; Möbius, B. Developments and paradigms in intonationresearch. *Speech Commun.* **2001**, *33*, 263–296. [CrossRef]
36. Tadeusiewicz, R. *Signal of Speech*; Publishing House of Communications: Warsaw, Poland, 1988.
37. Sawicka, I. Phonology. In *Grammar of Contemporary Polish. Phonetics and Phonology*; Wrobel, H., Ed.; "Od Nowa" Publishing House: Cracow, Poland, 1995; pp. 105–195.
38. Dłuska, M. *Prosody of the Polish Language*; PWN: Warsaw, Poland, 1976.
39. Pruszewicz, A. On the classification of voice quality and substitute speech in laryngectomized patients. *Otolaryngologia Polska* **1975**, *29*, 487–491. [PubMed]
40. Geertsema, A.A.; De Vries, M.P.; Schutte, H.K.; Lubbers, J.; Verkerke, G.J. In vitro measurements of aerodynamic characteristics of an improved tracheostoma valve for laryngectomees. *Eur. Arch. Otorhinolaryngol.* **1998**, *255*, 5, 244–249. [CrossRef]
41. Hook, J.; Noroozi, F.; Toygar, O.; Anbarjafari, G. Automatic speech based emotion recognition using paralinguistics features. *Bull. Pol. Acad. Sci. Tech. Sci.* **2019**, *67*, 3. [CrossRef]
42. Mik, Ł.; Lorenc, A.; Król, D.; Wielgat, R.; Święciński, R.; Jędryka, R. Fusing the electromagnetic articulograph, high-speed video cameras and a 16-channel microphone array for speech analysis. *Bull. Pol. Acad. Sci. Tech. Sci.* **2018**, *66*, 2018. [CrossRef]
43. Freitas, J.; Teixeira, A.; Dias, M.S.; Silva, A. An Introduction to Silent Speech Interfaces. In *SpringerBriefs in Speech Technology*; Springer: Berlin/Heidelberg, Germany, 2017; ISBN 978-3-319-40173-7.
44. Denby, B.; Csapó, T.G.; Wand, M. Future Speech Interfaces with Sensors and Machine Intelligence. *Sensors* **2023**, *23*, 1971. [CrossRef]
45. Wand, M.; Himmelsbach, A.; Heistermann, T.; Janke, M.; Schultz, T. Artifact removal algorithm for an EMG-based Silent Speech Interface. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013. [CrossRef]
46. Fagan, M.J.; Ell, S.R.; Gilbert, J.M.; Sarrazin, E.; Chapman, P.M. Development of a (silent) speech recognition system for patients following laryngectomy. *Med. Eng. Phys.* **2008**, *30*, 419–425. [CrossRef] [PubMed]
47. Gonzales, M.G.; Backer, K.C.; Yan, Y.; Miller, L.M.; Bortfeld, H.; Shahin, A.J. Audition controls the flow of visual time during multisensory perception. *iScience* **2022**, *25*, 104671. [CrossRef] [PubMed]
48. Gonzalez-Lopez, J.A.; Gomez-Alanis, A.; MartínDoñas, J.M.; Pérez-Córdoba, J.L.; Gomez, A.M. Silent Speech Interfaces for Speech Restoration: A Review. *IEEE Access* **2020**, *8*, 177995–178021. [CrossRef]

49. Gonzalez, J.A.; Cheah, L.A.; Gilbert, J.M.; Bai, J.; Ell, S.R.; Green, D.; Moore, R.K. A silent speech system based on permanent magnet articulography and directsynthesis. *Comput. Speech Lang.* **2016**, *39*, 67–87. [CrossRef]
50. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M.; Brumberg, J.S. Silent speech interfaces. *Speech Commun.* **2010**, *52*, 270–287. [CrossRef]