Vowel Recognition of Patients after Total Laryngectomy using Mel Frequency Cepstral Coefficients and Mouth Contour

Rafal W. Pietruch and Antoni D. Grzanka

Abstract—The paper addresses a problem of isolated vowels recognition in patients following total laryngectomy. The visual and acoustic speech modalities were separately incorporated in the machine learning algorithms. The authors used the Mel Frequency Cepstral Coefficients as acoustic descriptors of a speech signal. A lip contour was extracted from a video signal of the speaking faces using OpenCV software library. In a vowels recognition procedure the three types of classifiers were used for comparison purposes: Artificial Neural Networks, Support Vector Machines and Naive Bayes. The highest recognition rate was evaluated using Support Vector Machines. For a group of the laryngectomees having a different quality of speech the authors achieved 75% for acoustic and 40% for visual recognition performances. The authors obtained higher recognition rate than in a previous research where 10 cross-sectional areas of a vocal tract were estimated. Using presented image processing algorithm the visual features can be extracted automatically from a video signal.

Index Terms—Audio-visual speech recognition, Mel frequency cepstral coefficients, Neural networks, Face detection, Lipreading.

I. INTRODUCTION

A. Speech of laryngectomees

Laryngectomy is a partial or complete removal of a larynx. It is usually performed as a treatment for laryngeal carcinoma. While a partial laryngectomy often preserves the vocal folds, the patients following total laryngectomy lost the natural source of phonation. Thus, the most significant problem for them is to pronounce the vocalized sounds. Following the loss of the vocal cords patients are not able to phonate adequately. Their voice is hoarse, weak and stigmatized what makes it very difficult to understand by listeners. Moreover, the surgical procedure consequences include: loss of nasal function, changes of lung function, poor cough, swallowing difficulties and tracheostomal complications [1]. The main goal of a phoniatric rehabilitation is to learn how to articulate the alternative sounds. In oesophageal (OE) speech the pharyngo-esophageal segment (a zone located between the pharynx and esophagus) is used for voicing. The stomach and distal esophagus act as an air reservoir while cervical esophagus becomes a neoglottis [1]. A tracheoesophageal speech (TE) is an another surgical voice restoration. It involves the diversion of expired pulmonary air via a one-way valve [2]. The authors observed that certain percentage of the laryngectomees in Poland did not acquire an alternative voice. Many of the patients taking part in the research communicate with the silently articulated words called pseudo-whisper (PW). The pseudo-whisper of laryngectomees differs from usual whisper in natural speech because it cannot be generated using the air expired from lungs. The patients breath using tracheostoma, an opening through the neck into the trachea. Unless the surgical voice restoration the digestive and airway tracts are completely separated and the patient’s trachea is sewn to the skin, creating a tracheostoma. It should be pointed that most of the patients which took part in our experiment did not acquire alaryngeal speech. Even some of esophageal speakers were not able to phonate all vowels correctly.

B. Multimodal speech signal analysis

The main goal of the research is to enhance the speech to make it more intelligible for listeners. There are also several solutions for speech signal processing for the purpose of voice quality enhancement. In [3] authors developed special-purpose DSP hardware unit which replaces voicing sources of esophageal speech using the formant analysis-synthesis approach. The authors of [4] used digital signal processing technique based on modeling radiated pulses in frequency domain.

One of the solution is to create the automatic speech recognition (ASR) system and then artificial but more intelligible voice synthesis. The first two formants frequencies were often suggested as the most important factors of vowels recognition [5]. However, limiting the feature space to just two formants is not the reliable approach for automatic recognition of isolated vowels in pseudo-whisper [6]. There was no regular oscillation of the F1 and F2 formants observed. Several issues make evaluation of the classical acoustical descriptors of laryngectomees’ speech difficult. Noise from tracheostoma plays significant role in masking speech spectrum [6]. The effective length of vocal tract after laryngectomy is shortened and the formant frequencies are shifted to higher frequencies [2], [6]–[8].

In previous research [9] the authors used vocal tract cross-sectional areas calculated from linear prediction coefficients (LPC). For sample rate 8kHz the authors used the model of vocal tract divided into 10 resonance cavities. For these
acoustical features the authors achieved even lower recognition rate than for F1 and F2 formants. The neural network classifier was over-fitted by the training samples. Moreover, obtained vocal tract parameters were weakly correlated with visual factors including mouth area and jaw angle. The assumed speech production model was too simple and did not cover all sound generation details. In this research the vowels acoustic recognition is based on Mel-Frequency Cepstral Coefficients (MFCCs). This speech recognition approach was already applied for laryngectomees [10]. As it was shown the MFCC parameters are also capable for all vowels [11] and whole words [12] recognition. Therefore, the method can be used by authors in the future work.

Any supplementary information is essential for vowels recognition in pathological speech. Several multimodal approaches have been used by other authors for laryngectomees’ speech recognition. The example application with the magnets placed on the lips, teeth and tongue was presented in [13]. Miniature device with magnetic sensors reads the speech according to vocabulary database prepared for each patient. In this paper we incorporated additional visual features derived from face expression of the speaker. Most of applications combining the visual and auditory modalities are applied to noisy environment [14]–[19]. Similar method can be proposed for the recognition of low quality speech like pseudo-whisper. The authors of [20] achieved the 50% recognition performance using mouth contour (lip outline) information. With the audio quality degraded to -6dB, after incorporating visual features (with the 30% vision to sound weighting) the error rates drops from 27% to 16%. The problem of improving laryngectomees’ vowels recognition is comparable. Our research showed that the video data is a promising candidate for supporting laryngectomees’ speech analysis [21]. In [22] authors presented neural networks for audio-visual vowel recognition. It was shown that face expression doesn’t change after laryngectomy [9].

The authors presented an algorithm for face elements detection and mouth contour extraction from video signal [23]. It is a problem to get high quality data for this kind of studies. Unfortunately, the implemented method was sensitive to differences of face view like glasses and beards. Thus, in the previous research [9] the authors extracted the visual features manually inspecting related video frames. The visual features consisted of jaw angle and mouth area estimated from height and width of the mouth opening. Unfortunately, two pairs of vowels could not be separated according to extracted visual features (/æ/ from /ʌ/ and /i/ from /ɪ/). Therefore, the recognition of 4 classes of vowels was performed. The authors obtained following performance of vowels recognition: 43% for acoustical signal and 69% for visual features (classification limited to 4 classes).

C. Aims

The aim of the research was to compare authors methods presented in [9] with other solutions to audio-video speech recognition. The aim of the research was to assess the differences in vowels recognition performance between standard widely used methods and authors algorithms presented lately. A reason for changing the methods was the low recognition rates of those presented before. An algorithm which can be used in different environment on the different faces and voice should be presented. In the previous research the recognition of 4 classes of vowels was performed while in current research the authors distinguished all 6 Polish vowels using visual information. Thus, a method of two vowels pairs distinction should be presented (/æ/ from /ɛ/ and /i/ from /ɨ/). The another aim was to make the image processing fully automatic. The crucial aim of our work is to develop the mobile device for real time laryngectomees’ speech recognition and enhancement. The device will be equipped with the microphone and miniature video camera that will track the movement of face elements while speaking.

II. Methods

A. Subjects

In the research ten patients: 4 esophageal (3 males), one tracheoesophageal (male) and 5 pseudo-whisper speakers (3 males) were recorded. Three subjects (esophageal male, esophageal female and pseudo-whisper female) pronounced the entire material twice. Thus, every vowel was pronounced 13 times. For comparison purposes the patients were grouped according to the voicing method: oesophageal (OE), tracheoesophageal (TE) and pseudo-whisper (PW). As a material we used pronunciation of 6 isolated Polish vowels ‘a’, ’i’, ’e’, ’y’, ’o’, ’u’ (see appendix I for related IPA symbols). The words and short sentences were chosen for the future investigation of consonants analysis and recognition.

The video recording procedures restrictions included white background and visibility of entire face. The recordings were made in the small (about 3m²) room with sound proof doors and small window to operator’s room.

B. Signal Acquisition

The recordings were made with the use of Panasonic NV-D565EG video camera. An external microphone Rode VideoMic was connected to the camera in the 20cm distance. Both devices were at the same distance to the patient face: about 1.5m. The recordings were copied from DV tape to the files in planar YUV 4:2:0 format (one blue- and one red-differences chroma components per 2x2 luma samples). The video image resolution was 720x576 pixels and the sample rate was 25 fps (frames per second). Monophonic acoustic signal was sampled with 48000 Hz and stored in PCM 16bit format. Acoustic signal was manually divided to the fragments related to each isolated vowel using FFMpeg program. The acoustic signal of each vowel was normalized to (−1, 1) range. Each fragment was limited to an articulation of the single vowel: the silence before and after articulation was removed from the signal. Values in the interval between (−0.4, 0.4) were considered as silence. Each sample was divided to analysis windows with the same length. Thus, the number of tokens for every subjects depended on the time of the vowel articulation. The visual features were extracted from the video frames related to the time of acoustic parameters extraction. Video
Frames from every vowel articulation fragment were dropped to the picture images of with lossless data compression (PNG format). Every image was processed for visual features extraction. The visual inspection of processed image was made to check from which frames the parameters were successfully extracted. Frames where the image processing failed were rejected from the analysis. Thus, the number of visual data differs for each vowel, and so on from subject to subject.

C. Acoustic model

Eleven MFCC coefficients were extracted from 48kHz acoustic signal using R program (version 2.11.1) and 'tuneR' package (version 0.3-0) [24]. The MFCC parameters computation is performed in five steps [24]. The acoustic signal is pre-emphasized with the finite impulse response (FIR) high-pass filter with transfer function \( H(z) = 1 - a z^{-1} \), where coefficient \( a = 0.1 \). The signal is converted to frames of length 0.025 seconds using a Hamming window to avoid negative effects on the edges of each frame. A discrete Fourier transformation (DFT) is then computed. The signal is mapped to the Mel scale filter bank consisting of 24 triangular filters, which overlap by a factor 0.25. In the last step an N-point \((N = 11)\) inverse discrete cosine transformation is applied to the signal (see equation 1).

\[
MFCC_i = \sum_{n=1}^{N} x_n \cos[i(n - \frac{1}{2}) \frac{\pi}{N}]; \quad i = 1, 2, \cdots, N
\]  

D. Visual features

The authors used OpenCV library [25] (version 2.2) for face, eyes and mouth detection from video samples. The Haar cascade classifiers introduced by Viola and Jones [26] were used according to [27]. We used an algorithm of snake contour which changes its shape to minimize an energy [28]. In this case a binary image of mouth region was treated as a energy. The snake algorithm was also implemented in the OpenCV library. The distances between arithmetic center of mouth border and thirty contour points formed the input data for visual classifier. Each distance was normalized to the distance between the eyes.

E. Classification

To obtain vowels classification from visual and acoustic modalities three kinds of statistical machine learning algorithms were used: artificial neural network (ANN), Support Vector Machines (SVM), and Naive Bayesian classifier. The authors used single hidden layer feed-forward neural network. The optimization was performed via the BFGS method [29]. A Log-Sigmoid transfer function for the first layer and linear transfer function for output layer were chosen. In all cases hidden layer size was 5. Initial random weights were chosen from \([-0.5, 0.5]\) range. The weight decay was set to \(5 \times 10^{-4}\). Maximum number of iteration was 600. The ANN was trained to assign maximum value \((1)\) for output related to spoken vowel and minimum value \((0)\) to other outputs. In the ANN simulation the output of maximum value indicated recognized vowel. In SVM classifier a radial kernel was used \(exp(-\gamma|u-v|^2)\) [30] with \(\gamma\) parameter equal 1/(data dimension).

Every classifier was trained to recognize 6 Polish vowels ‘a’, ‘i’, ‘e’, ‘y’, ‘o’, ‘u’. Before the training process the input data was multiplied to have the same number of tokens in every vowels group. The training group consisted of randomly chosen half of each vowel samples. The remaining half of samples formed the classifier validation group. Acoustic and visual features were analyzed separately.

All statistical analyzes were implemented in the R program. We used nnet package [31] for neural networks and e1071 [32] for support vector machine and naive Bayes classifiers.

III. RESULTS

A. Image processing results

The example results of the image processing algorithm are shown on the following figures. Figure 1(a) is the input image for processing and on the figure 1(b) one can see the part of the image converted to gray color. First the gray image is masked using the mouth region retrieved using Haar cascade classifier. After the smoothing and normalization of the masked region the result is shown on figure 1(c). Then a binary threshold is applied, and the result is shown on the figure 1(d). The detected contours can be seen on figure 1(e). Finally, the result of snake algorithm and detected eyes regions are shown on the figure 1(e).

B. Classification results

The number of tokens obtained from acoustic and visual modalities are given in table I.
The recognition percentage according to different modalities, training and simulation groups are given in Table II.

It classified training data with higher performance while recognition rate of validation samples was lower like for MFCC parameters all subjects. Thus, the better solution seem to be using SVM classifier for limited samples.

The recognition rate of visual descriptors is lower than reported in literature [20]. The recognition rate of speech using mouth contour estimation failed. The extension of visual features are weakly correlated with the acoustic model presented in [9] which have intuitive equivalents in visual features from vowel articulation time divided to many fragments.

Compared with previous research the authors achieved higher recognition rate of alaryngeal speech using healthy speech was rejected in this paper. The authors limited the recognition experiments for laryngectomees group. It was possible because more training tokens were evaluated compared with the previous research. To increase the amount of samples the authors measured the acoustic and visual features from vocal articulation time divided to many fragments of the same length.

In [22] acoustic and visual parameters were both limited to 2 dimensions. Relying on only two formants when analyzing vowel recordings was not sufficient anyway. Moreover, the F1 and F2 formants were extracted from speech spectrum manually by the operator. The a priori information about possible formant positions for vowels were taken into consideration. In the previous research the authors did not use the samples of pseudo-whisper voice because of problems with acoustic features evaluation reported in [6]. Moreover, an effective length of the vocal track after laryngectomy is shortened and the formants are shifted to higher frequencies. The differences in average formant frequencies between post laryngectomy and natural voice yielded significant error in vowels recognition from acoustic features. In presented paper both alaryngeal and pseudo-whisper groups were taken for training and testing. The authors achieved higher recognition rate of alaryngeal speech based on MFCC parameters (98%) than from F1-F2 formants (75%) [22]. The same percentage 98% was achieved in [22] for control group in training task. Obtaining 70% recognition rate of esophageal speech. No comparison could be also made about the difference of the recognition performance between the control group and the esophageal group. To design a dedicated system for esophageal speech training the ASR using healthy speech was rejected in this paper. The authors limited the recognition experiments for laryngectomees group.

In a part of experiment the authors used healthy speech for training and esophageal speech for testing. This scenario disabled the representative results and yielded poor recognition rate of esophageal speech. No comparison could be also made about the difference of the recognition performance between the control group and the esophageal group. To design a dedicated system for esophageal speech training the ASR using healthy speech was rejected in this paper. The authors limited the recognition experiments for laryngectomees group. It was possible because more training tokens were evaluated compared with the previous research. To increase the amount of samples the authors measured the acoustic and visual features from vocal articulation time divided to many fragments of the same length.

In the previous research [9] the same classifier for a control group and the patients was applied because of limited samples. In a part of experiment the authors used healthy speech for training and esophageal speech for testing. This scenario disabled the representative results and yielded poor recognition rate of esophageal speech. No comparison could be also made about the difference of the recognition performance between the control group and the esophageal group. To design a dedicated system for esophageal speech training the ASR using healthy speech was rejected in this paper. The authors limited the recognition experiments for laryngectomees group. It was possible because more training tokens were evaluated compared with the previous research. To increase the amount of samples the authors measured the acoustic and visual features from vocal articulation time divided to many fragments of the same length.

In [22] acoustic and visual parameters were both limited to 2 dimensions. Relying on only two formants when analyzing vowel recordings was not sufficient anyway. Moreover, the F1 and F2 formants were extracted from speech spectrum manually by the operator. The a priori information about possible formant positions for vowels were taken into consideration. In the previous research the authors did not use the samples of pseudo-whisper voice because of problems with acoustic features evaluation reported in [6]. Moreover, an effective length of the vocal track after laryngectomy is shortened and the formants are shifted to higher frequencies. The differences in average formant frequencies between post laryngectomy and natural voice yielded significant error in vowels recognition from acoustic features. In presented paper both alaryngeal and pseudo-whisper groups were taken for training and testing. The authors achieved higher recognition rate of alaryngeal speech based on MFCC parameters (98%) than from F1-F2 formants (75%) [22]. The same percentage 98% was achieved in [22] for control group in training task. Obtaining 70% recognition performance for pseudo-whisper vowels the authors achieved an advantageous result compared with the previous research.

The authors resigned from using cross-sectional areas presented in [9] which have intuitive equivalents in visual features. Unfortunately it was also presented that proposed visual features are weakly correlated with the acoustic model parameters. It was already known [33] that cross-sectional areas calculated for a given speech wave are ambiguous. The low recognition rate (43%) was also a reason for changing the

### Table I

<table>
<thead>
<tr>
<th>modality</th>
<th>group</th>
<th>/a/</th>
<th>/e/</th>
<th>/i/</th>
<th>/o/</th>
<th>/u/</th>
<th>/y/</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>OE, TE</td>
<td>141</td>
<td>71</td>
<td>92</td>
<td>96</td>
<td>114</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>PW</td>
<td>193</td>
<td>152</td>
<td>178</td>
<td>199</td>
<td>198</td>
<td>156</td>
</tr>
<tr>
<td>M. Contour</td>
<td>OE, TE, PW</td>
<td>85</td>
<td>119</td>
<td>135</td>
<td>125</td>
<td>116</td>
<td>122</td>
</tr>
</tbody>
</table>

### Table II

Sustained vowels recognition rates of acoustic (MFCC) and visual (mouth contour) modalities. Results obtained for all patients, alaryngeal voice (oesophageal and tracheoesophageal) and for pseudo-whisper (PW) group.

<table>
<thead>
<tr>
<th>input</th>
<th>group</th>
<th>classifier</th>
<th>training</th>
<th>validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>OE,TE,PW</td>
<td>SVM</td>
<td>91%</td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N. Bayes</td>
<td>60%</td>
<td>51%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN</td>
<td>75%</td>
<td>57%</td>
</tr>
<tr>
<td></td>
<td>OE,TE</td>
<td>SVM</td>
<td>100%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N. Bayes</td>
<td>92%</td>
<td>91%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN</td>
<td>100%</td>
<td>93%</td>
</tr>
<tr>
<td></td>
<td>PW</td>
<td>SVM</td>
<td>85%</td>
<td>70%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N. Bayes</td>
<td>52%</td>
<td>45%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN</td>
<td>73%</td>
<td>49%</td>
</tr>
<tr>
<td>Mouth Contour</td>
<td>OE,TE,PW</td>
<td>SVM</td>
<td>81%</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N. Bayes</td>
<td>40%</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANN</td>
<td>79%</td>
<td>37%</td>
</tr>
</tbody>
</table>

### Table III

Validation result of MFCC based recognition of laryngectomees vowels.

<table>
<thead>
<tr>
<th>classified vowel class</th>
<th>/a/</th>
<th>/e/</th>
<th>/i/</th>
<th>/o/</th>
<th>/u/</th>
<th>/y/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>108</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>/e/</td>
<td>7</td>
<td>103</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>/i/</td>
<td>9</td>
<td>12</td>
<td>87</td>
<td>1</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>/o/</td>
<td>2</td>
<td>9</td>
<td>1</td>
<td>98</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>/u/</td>
<td>4</td>
<td>12</td>
<td>8</td>
<td>13</td>
<td>94</td>
<td>12</td>
</tr>
<tr>
<td>/y/</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>104</td>
</tr>
</tbody>
</table>

### Table IV

Validation result of mouth contour based recognition of laryngectomees vowels.

<table>
<thead>
<tr>
<th>classified vowel class</th>
<th>/a/</th>
<th>/e/</th>
<th>/i/</th>
<th>/o/</th>
<th>/u/</th>
<th>/y/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>46</td>
<td>20</td>
<td>12</td>
<td>4</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>/e/</td>
<td>12</td>
<td>42</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>/i/</td>
<td>11</td>
<td>12</td>
<td>28</td>
<td>6</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>/o/</td>
<td>4</td>
<td>2</td>
<td>25</td>
<td>79</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>/u/</td>
<td>2</td>
<td>6</td>
<td>25</td>
<td>13</td>
<td>79</td>
<td>116</td>
</tr>
<tr>
<td>/y/</td>
<td>19</td>
<td>6</td>
<td>22</td>
<td>7</td>
<td>2</td>
<td>29</td>
</tr>
</tbody>
</table>

IV. Discussion

In the previous research [9] the same classifier for a control group and the patients was applied because of limited samples. In a part of experiment the authors used healthy speech for training and esophageal speech for testing. This scenario disabled the representative results and yielded poor recognition rate of esophageal speech. No comparison could be also made about the difference of the recognition performance between the control group and the esophageal group. To design a dedicated system for esophageal speech training the ASR using healthy speech was rejected in this paper. The authors limited the recognition experiments for laryngectomees group. It was possible because more training tokens were evaluated compared with the previous research. To increase the amount of samples the authors measured the acoustic and visual features from vocal articulation time divided to many fragments of the same length.

In [22] acoustic and visual parameters were both limited to 2 dimensions. Relying on only two formants when analyzing vowel recordings was not sufficient anyway. Moreover, the F1 and F2 formants were extracted from speech spectrum manually by the operator. The a priori information about possible formant positions for vowels were taken into consideration. In the previous research the authors did not use the samples of pseudo-whisper voice because of problems with acoustic features evaluation reported in [6]. Moreover, an effective length of the vocal track after laryngectomy is shortened and the formants are shifted to higher frequencies. The differences in average formant frequencies between post laryngectomy and natural voice yielded significant error in vowels recognition from acoustic features. In presented paper both alaryngeal and pseudo-whisper groups were taken for training and testing. The authors achieved higher recognition rate of alaryngeal speech based on MFCC parameters (98%) than from F1-F2 formants (75%) [22]. The same percentage 98% was achieved in [22] for control group in training task. Obtaining 70% recognition performance for pseudo-whisper vowels the authors achieved an advantageous result compared with the previous research.

The authors resigned from using cross-sectional areas presented in [9] which have intuitive equivalents in visual features. Unfortunately it was also presented that proposed visual features are weakly correlated with the acoustic model parameters. It was already known [33] that cross-sectional areas calculated for a given speech wave are ambiguous. The low recognition rate (43%) was also a reason for changing the
acoustic signal processing methods. Recognition performance based on 10 cross-sectional areas [9] is even lower compared with formant-based classifier [22]. Using MFCC parameters applied to 48000Hz acoustic signal higher recognition performance was obtained compared to F1-F2 formants and cross sectional areas calculated from 8kHz speech. Moreover, proposing MFCC as the acoustic features the authors overcame over-fitting of classifier [9]. The recognition rate based on MFCC parameters for pseudo-whisper group is still lower than for alaryngeal patients (OE, TE) (see table II). However, the difference is not so evident comparing to previous research [6].

It was presented [22] that in pseudo-whisper an acoustic vowel recognition process can be supplied by visual factors. Because the recognition procedure based on F1-F2 formants failed for pseudo-whisper group [6] the authors proposed a fusion of acoustic and visual modalities suitable for patients voice analysis. However, in [22] two pairs of vowels could not be classified according to visual features. Thus, the authors could not compare the results of the acoustic 6 classes and visual 4 classes recognition tasks. It was also not advantageous to combine two modalities in all vowels classification task. With the use of mouth contour the authors were able to distinguish these two pairs of vowels as presented in table IV. However, according to the presented results visual and acoustic factors fusion would not enhance the recognition performance. The authors achieved much lower recognition performance of visual factors (40%) compared with MFCC parameters even for pseudo-whisper voice (70%).

In the visual signals another problem was related to the head position of subjects, many of recordings were rejected because of unusable data. Moreover, a human body presents high interpersonal variability [34]. There were situations where the mouth was not visible enough to measure usable parameters. The performance of algorithm was enhanced with the use of uncompressed video files and by changing light conditions of video recordings. With the use of Haar cascades the detection of head and face elements is more general and more reliable. However, there was a problem with the opened mouth detection. Moreover, in many images the border of lips can be elusive [34] According to [15], it would be also advantageous to include first derivative of the visual features.

V. CONCLUSIONS

Compared with the previous research [9] we achieved several enhancements in audio-visual voice recognition. The MFCC parameters are suitable factors for acoustic vowel recognition of laryngectomees. This method is also capable for the pseudo-whisper voice. The presented algorithm of visual face expression analysis can distinguish two vowels pairs: /a/ from /e/ and /i/ from /y/. The visual factors obtained using the image processing method can be used for statistical analysis and vowels classification. The only manual procedure was the elimination of incorrectly processed images. However, the performance of vowels recognition based on mouth contour is about 10% less than presented in literature [20]. On this stage combining acoustic and visual modalities would not be advantageous.

VI. FUTURE WORK

The presented algorithms should be tested in other conditions. The authors should use another camera (e.g. Internet camera) in different environment (indoor and outdoor) to achieve different color and light conditions. It might be worth to test the presented methods with wider portion of the earth’s population (especially different skin colors). Then the authors should generalize the algorithms to be suitable for different recordings conditions. There can be used skin detection approach [35] which are expected to result in high performance. Some people wear glasses and beards which makes face elements detection difficult [23]. In the presented research there were no subjects with beards or glasses. However, the authors should be aware of that problems when enhancing the presented algorithms. The algorithm of face elements detection should be extended with a Haar cascade for an open mouth. The heuristic settings of visual algorithms constants like thresholds or snake function parameters should be eliminated. The authors should modify the computer vision algorithm to achieve similar or higher recognition performance than in [20]. There can be used another methods to track lips contour like dynamic contours presented in [20]. Teeth visibility could also enhance the recognition performance, example application using neural network was presented in [36]. Thirty control points used it this research can be reduced to smaller dimension using e.g. PCA. The results could be then compared with the literature [20]. In the future work a recognition performance of consonants and words [12] should be evaluated. The authors are going to develop a mobile devices equipped with a microphone and digital video camera that will incorporate the presented methods.

VII. ACKNOWLEDGMENT

The authors would like to thank prof. Wieslaw Konopka from Medical University of Lodz (Poland) for organizing the recording sessions.

APPENDIX I

IPA SYMBOLS FOR POLISH VOWELS.

Polish vowels transcriptions are given in table V.

REFERENCES


