Multimed Tools Appl (2018) 77:16495–16532
https://doi.org/10.1007/s11042-017-5217-5

A comparative study of English viseme recognition methods and algorithms

Dawid Jachimski[1] · **Andrzej Czyzewski**[1] · **Tomasz Ciszewski**[1]

Received: 4 February 2017 / Revised: 18 August 2017 / Accepted: 8 September 2017 /
Published online: 7 October 2017

Abstract An elementary visual unit – the viseme is concerned in the paper in the context of preparing the feature vector as a main visual input component of Audio-Visual Speech Recognition systems. The aim of the presented research is a review of various approaches to the problem, the implementation of algorithms proposed in the literature and a comparative research on their effectiveness. In the course of the study an optimal feature vector construction and an appropriate selection of the classifier were sought. The experimental research was conducted on the basis of a spoken corpus in which speech was represented both acoustically and visually. The extracted features represented three types: geometrical, textural and mixed ones. The features were processed employing the classification algorithms based on Hidden Markov Models and Sequential Minimal Optimization. Tests were carried out employing the processed video material recorded with English native speakers who read specially prepared list of commands. The obtained results are discussed in the paper.

Keywords viseme · Parameterization of mouth region · Support Vector Machine · Hidden Markov Model · Pattern recognition · Audiovisual speech recognition

1 Introduction

The methods of algorithmic viseme recognition have been developed and discussed in the literature for a relatively long time. Despite the progress in the area, however, they still do not produce fully satisfactory results in the recognition of speech elements on the basis of lip picture (viseme) analysis. The problem of automatic viseme recognition is closely related to

✉ Andrzej Czyzewski
ac@pg.gda.pl

[1] Multimedia Systems Department, Gdańsk University of Technology ETI Faculty, ul. Narutowicza 11/12, Gdańsk, Poland

research on automatic speech recognition which was initiated in mid 20th century, e.g. in the proposal of an audio-visual speech recognition system (AVSR) by Petajan et al. [27]. The processing of an additional set of visual data may enable the extraction of information leading to the recognition enhancement of linguistic units. The analysis of visual signals may concentrate on units such as phonemes and visemes, isolated words, sequences of words and continuous/spontaneous speech. The viseme is a visual counterpart of the phoneme [7].

The signature of a viseme is a particular picture frame, i.e. a static image of the speaker's face. There also exists another, less popular definition, according to which visemes may be understood as articulatory gestures, lip movement, lip position, jaw movement, teeth exposition, etc. [2]. Certain phonemes may have the same visual representation [3, 20, 25]. What follows is that it is not a one-to-one relation. A given facial image may, thus, be identical for different realizations of the same phoneme depending on its phonetic environment. Therefore, preliminary classification (division) is necessary. Relying entirely on the visual input may lead to the erroneous classification of an utterance, e.g. "*elephant juice*" may be recognized as "*I love you*" [41]. It has also been shown that the deprivation of the visual input has a detrimental effect on human perception and leads to lower (by 4 dB) tolerance of noise in the acoustic environment [13].

In the present study an approach is proposed which is based on the analysis of visemes. Phones were first classified into the corresponding phonemes and then the phonemes were assigned to appropriate classes of visemes. A selection of commands in English (recorded as a linguistic corpus) was recorded audio-visually by a group of native speakers of English. The material prepared in Gdansk University of Technology has also been made available to research community in the form of a multimodal database accessible at the address: http://www.modality-corpus.org/.

Section 2 which follows this introduction presents theoretical methods of viseme classification. It contains also a description of the phoneme-to-viseme map used for the research. Section 3 describes the algorithms employed to the automatic detection of ROI of the mouth followed by feature extraction and classification methods presentation. The experimental setup configuration and data preparation are discussed in Sections 4 and 5, whereas in Section 6 obtained results were arranged in a comparative manner. The last section refers to conclusions and directions for further research.

2 Viseme classification methods

According to the basic definition, the viseme is the smallest recognizable unit correlated with a particular realization of a given phoneme. This definition, however, does not determine the ways in which visemes can be classified into groups. The precise number of all possible visemes, which may depend on the assumed classification criteria, is not provided. The number of visemes may oscillate between a dozen and a few thousands. The most popular classifications confine the set of visemes to approximately 10–20 groups.

There are two major criteria of classifying visemes [2]:

- according to the facial image, the shape and arrangement of the lips, teeth exposition during the articulation of particular linguistic units, and
- according to the phonemes with an identical visual representation.

The second definition is especially popular since it facilitates the preparation of training and testing data.

Drawing on precisely described phonemic models substantially reduces the amount of work. By analogy, some of the results of an earlier research on acoustic speech recognition can be utilized. However, there exist no reliable and unambiguous tests confirming that this is a better method. Undoubtedly, the advantage of this approach is analogy and viseme-phoneme correlation.

The second method facilitates the construction of the viseme <−> phoneme map. The map will be of the many-to-one type representation since thanks to this approach a few phonemes can have the same visual realization. The way in which this representation is constructed can be based on certain simplifications in the assumed classification method. The most popular methods are:

– *linguistic* – the classes of visemes are defined on the basis of an intuitive linguistic classification of groups of phonemes according to their expected visual realization,
– *data driven* – the classes of visemes are defined on the basis of data acquired through parameter extraction and clustering [40].

The data-driven method has a number of advantages over the purely theoretical linguistic approach. Speech processing systems are based on statistical models which are arrived at on the basis of data and not on the assumed results and structures. The linguistic method, on the other hand, facilitates a precise description of visemes included in a given linguistic unit. It may, however, turn out to be more imprecise as it relies on an intuitive approach. Considering the fact that as yet no generally accepted classification model has been proposed and the linguistic approach has not evolved into a standard mature model, the research on this issue may produce interesting results. The principle for carrying out the transcription of commands is illustrated Fig. 1.

In this work a model based on the most popular way of classifying visemes, i.e. MPEG-4 [36], has been assumed. It is the most important component determining the *Face Animation Parameters* marked out during face animation. The classification is based on the linguistic analysis of articulatory similarities of phonemes occurring in the commands used in the audio-visual material included in the database. The analysis takes into account the following articulatory features and assumptions:

– the exclusion of diphthongs since they are dynamic vowels and their imaging will include the component features if the starting point and the glide;

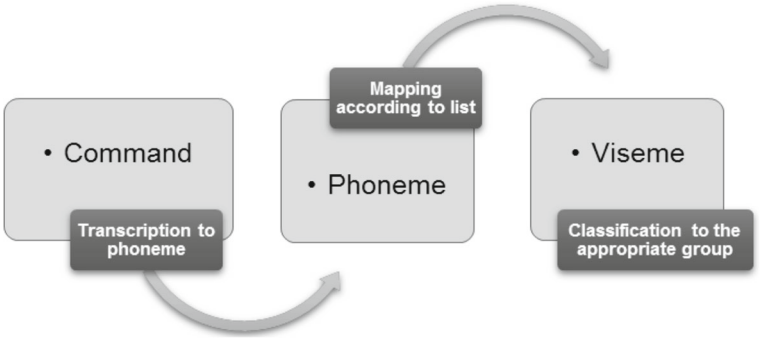


Fig. 1 Flowchart illustrating the principle of command transcription

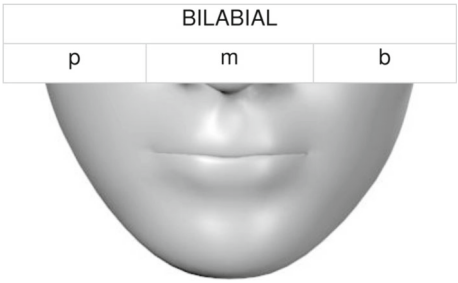


Fig. 2 Theoretical image of W1 group of visemes

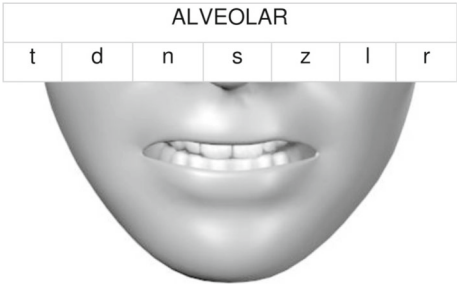


Fig. 3 Theoretical image of W2 group of visemes

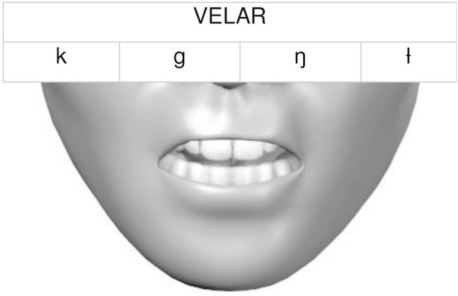


Fig. 4 Theoretical image of W3 group of visemes

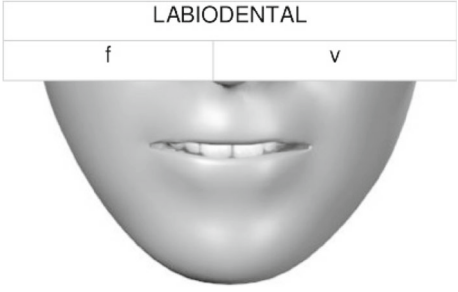


Fig. 5 Theoretical image of W4 group of visemes

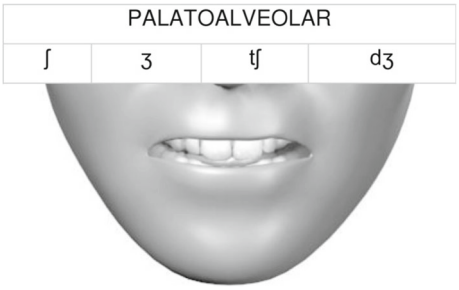


Fig. 6 Theoretical image of W5 group of visemes

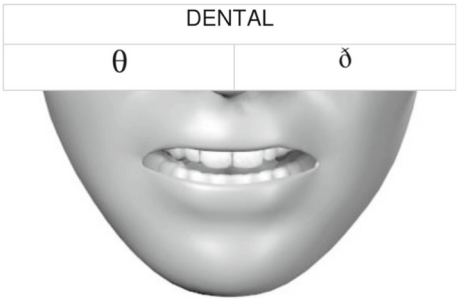


Fig. 7 Theoretical image of W6 group of visemes

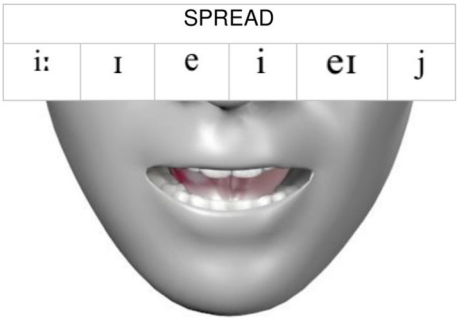


Fig. 8 Theoretical image of W7 group of visemes

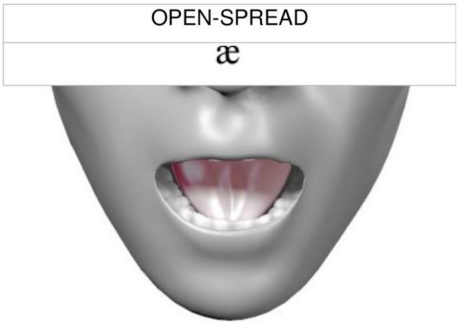


Fig. 9 Theoretical image of W8 group of visemes

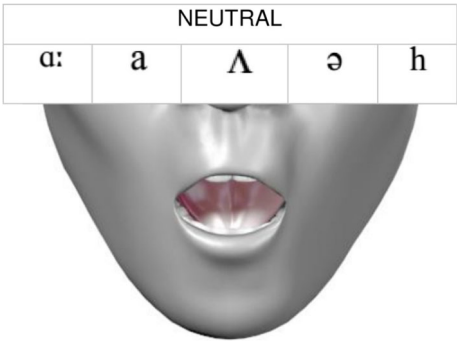


Fig. 10 Theoretical image of W9 group of visemes

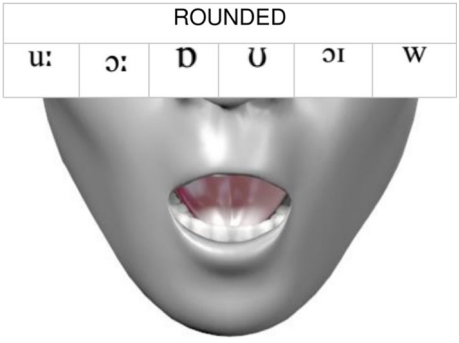


Fig. 11 Theoretical image of W10 group of visemes

Fig. 12 Theoretical image of W11 group of visemes

Fig. 13 Theoretical image of W12 group of visemes

– consonants assume the articulatory lip settings of the following vowels, i.e. /k/ in the words *keep* will have the features of the /i:/ vowel and /k/ in the word *cool* will have the features of the /u:/ vowel;
– 'dark' /l/ which is a velarized variant of the lateral consonant /l/ and occurs word-finally or before another consonant has the articulatory features which are identical with /k, g/ consonants;
– unobstructed consonants /h, j, w/ will have a 'vocalic' imaging, hence their inclusion in the vocalic table.

Our model contains 12 classes of visemes into which the relevant phonemes have been classified. In Figs. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 the theoretical shapes of the lips are presented which illustrate particular phonemes.

The phoneme-to-viseme mapping is shown in Table 1. It includes 6 classes of consonantal visemes and 5 classes of vocalic visemes. The silence viseme is an important element of the classification and has also been taken into account. The set of the most similar phonemes ascribed to particular classes is also included In Table 1. The resulting classification and the corresponding map are representative of the linguistic approach.

Fernandez-Lopez et al proposed viseme groups for the Spanish vocabulary using the phonemes with a similar appearance [8]. In our paper we have build the visemes groups based on a similar approach. The difference is however, that our research describes several types of parameters and it gathers the scores for diversified sets containing them.

In the literature there also appear proposals of other maps: *linguistic*, *linguistic-data driven* and *data-driven*. The sizes of particular classes also differ. An example of a map which includes a different number of viseme classes is found in Neti et al. classification used by IBM for constructing the *ViaVoice* viseme database employing three neighboring visemes and the MPEG-4 map [26].

The selection of an appropriate model is a difficult task, given the lack of comparative tests. There are few studies analyzing the results obtained for a particular model in the same testing environment and based on the same collection of data. However, such analyzes appear more often now which means that the need for developing viseme-based systems is becoming recognized as the right direction in AVSR research. The theoretical images for particular viseme groups are presented in Figs. 2–13. They were generated using the *Verbots Tools Conversive Character Studio Visemes* [32] available through an open-source license *GNU*.

3 Algorithms for detecting the location and shape of the lips in the image

The first task which enables further viseme analysis is the detection of the speakers' lip area. The extraction of information concerning the shape of the lips is carried out in a few steps.

The first step is the detection of the *Region of Interest* (ROI). The correct localization of the speaker's lips is of great importance for the effectiveness of algorithms which detect the key points in the face area [9, 15, 18].

The algorithms which detect the lip area are based on recognizing certain patterns which are standardized and widely used. They use the dependencies between the eyes, eyebrows, nose and lips. The detection of the face area and the application of an algorithm searching for similarities and dependencies in mutual localization of particular elements enables an effective recognition of ROI and the subsequent feature extraction [39]. This is a critical

Table 1 Classification of phonemes into groups belonging to a particular viseme

	Groups of visemes					
	CONSONANTAL					
	LABIAL	ALVEOLAR	VELAR	LABIODENTAL	PALATO-ALVEOLAR	DENTAL
PHONEMES	p	t	k	f	ʃ	θ
	b	d	g	v	ʒ	ð
	m	n	ŋ		tʃ	
		s	ɬ		dʒ	
		z				
		l				
		r				
	VOCALIC					Silence
	SPREAD	OPEN-SPREAD	NEUTRAL	ROUNDED	PROTRUDING-ROUNDED	CLOSED
	iː	æ	ɑː	uː	ɜː	#
	ɪ		a	ɔː		
	e		ʌ	ɒ		
	i		ə	ʊ		
	eɪ		h	ɔɪ		
	j			w		

element since the precise localization of lips in the facial image conditions the effectiveness of the following stages of analysis.

The systems which are developed usually make it possible to additionally localize and describe the lip contours [1]. Additionally, taking into account the shape of the lips enables a precise description of analyzed picture frames.

At the moment the most effective and most frequently used methods are based on *Active Appearance Models*. AAM is a domain of statistical models describing the appearance and shape of certain characteristic objects. The result of an algorithm application is to a generated universal description of particular objects. These models allow establish the set of characteristic points describing the features of an object. This approach was used in the experiments carried out and described in further parts of the paper. The schema of the implemented algorithm based on procedures known from the cited literature is presented in Fig. 14.

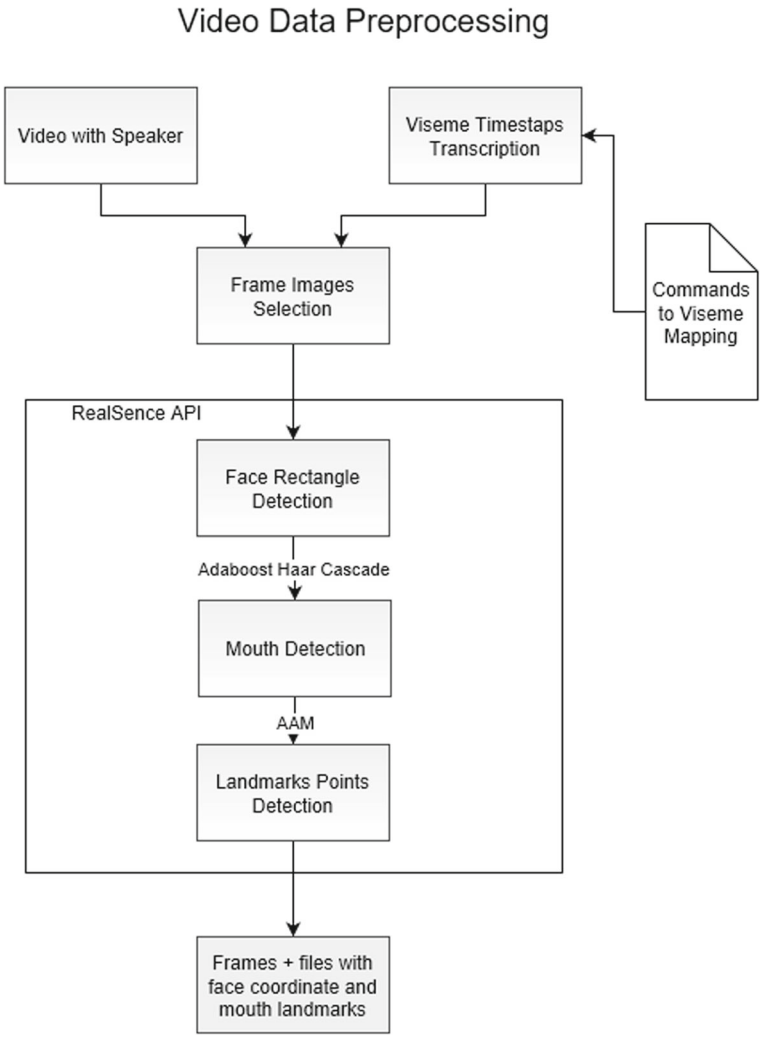


Fig. 14 Video data preprocessing algorithm schema

Thanks to the AAM a model can be created which transfers not only the information about the shape of an object but also about the distribution of pixel brightness in a frame, the color of particular component elements and their texture. Due to a wide range of analyzed parameters, this approach should be classified as the group of hybrid algorithms which are successfully and effectively used in many areas of research.

Generally, the shape of an object is defined by a set of points which are located in characteristic places of the object, placed at its edges or inside it. On the basis the points of the shape the representation of the shape of an object is determined. The Active Shape Model (ASM) [5] algorithms and its immediate successor, the Active Appearance Model [4, 30], are examples of this approach. Both algorithms use the same definition of the shape of an object but differ in their representation of the appearance of the object. In the ASM method for every point of the shape it is the appearance of an object in the proximity of the point, usually represented by a vector including the color, texture and the gradient of the image. The AAM method, on the other hand, includes all pixels of an object within its contour.

Statistical models based on the ROI analysis and the detection of outer and inner lip contours are the main ways of detecting and circumscribing them with key points (Point of Interest, POI). These points are used in the calculation of vector parameters differentiating particular lip setting in the articulation of phonemes [6].

Research on human perception clearly shows that lip-reading information is used in speech processing [12, 42]. Speech perception utilizes such elements as the visibility of upper/lower teeth and the degree of tongue visibility. It is vital then that the algorithms simulate such behavior. The information extracted from the visual input should therefore include such data.

The extraction of possibly biggest number of precise elements (the key points circumscribed on the analyzed object which create the model) is an important aspect of constructing automatic lip detection and marking systems. The first stage is marking the outer lip contours. Most algorithms are based on the analysis of brightness changes in the transition between the areas adjacent to the lips and the lips themselves. Then, the image is subject to a similar analysis carried out for the inner contours. Parameter extraction from the area inside the lips is more difficult and simultaneously more important for viseme recognition. Particular classes of visemes differ in terms of teeth and tongue visibility and the degree of their exposition in the picture frame [15].

Statistical models are built which include transition and similarity matrices for the speakers' lip shapes. The key problem is the selection of a model catering for the transitions between the dark and bright valleys during the analysis. The area inside the lips changes dynamically which makes it difficult to work out a universal hierarchy of the model. The algorithms deal with the problem by using statistical Bayes classifiers and Fisher linear distribution function [22, 34, 35].

In order to arrive at the model a training set must be prepared which includes a diversified selection of different lip images: closed, half-open, with visible/invisible teeth (in different variants) and visible/invisible tongue. Then, ideal initial threshold values are calculated. For the data obtained during the algorithm application the maximal membership probabilities of a given component are calculated for every pixel in the previously established area of interest [15]. These points are subject to clustering using *K-means* method. On the basis of the results a decision can be made whether a given pixel belongs to the set of contour points or not.

The models include information concerning the shape and the structure of a given facial image as well as additional information about their modification or possible changes of their shape. Usually, the models block the possibility of an incorrect realization of a particular

shape. Thus, it is possible to preserve the standard original face for real shapes and avoid the danger of unnatural mutations and deformations. For every picture provided as input for the algorithm in the detection phase comparisons are made between the shape and the database. At the moment when the highest probability value of feature vector matching of the currently analyzed frame with those included in the database is achieved, the classification decision is made [14].

An sample result of the algorithm application is presented in Fig. 15.

3.1 Preparation of visual feature parameter vector

The lips represented by key points determined with AAM algorithms cannot be directly used to represent the actual features extracted from the speaker's lips. Methods for calculating the parameter vector must be worked out in order to sub-divide particular lip arrangements. The initial analytical problem is also the fact that the parameter vector must be made independent of an individual speaker [27].

The localization of lip contours also entails other problems since it does not take into account the tongue and the teeth position in the picture frame. It is not uncommon that classifiers do not provide precise information regarding the position of particular elements in the lip area itself. The algorithms based on the principle of typical tonal distribution recognition in greyscale encounter problems as the contrast changes in picture frames [15]. Such complications call for devising a method of describing lips with parameters which will most robustly differentiate particular lip settings. Possible ways of representing those features will be discussed in further sections.

The detected lip contours can be represented by means of a rectangle which circumscribes them. Due to a potentially large number of pixels belonging to this area, the parameter vector may be too costly, both in terms of data storage and computation. The

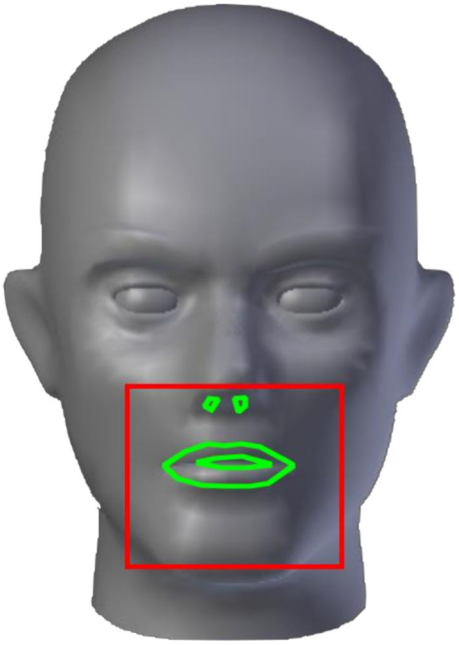


Fig. 15 A sample result of AAM algorithms application for the lip

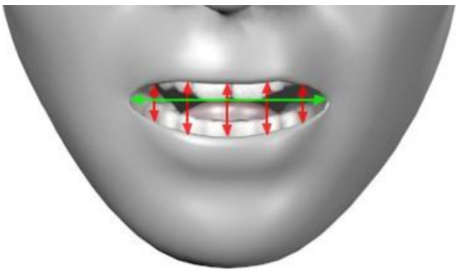


Fig. 16 Sample distances inside the ROI lip area

number of pixels (which often contains values for a few components) may reach as many as several hundreds of thousands. In order to reduce the dimensionality, and as a result the cost, *Discrete Cosine Transformation* (DCT) is used [29]. It is a typical transformation used for picture compression. Thanks to this transformation the multidimensionality of a vector can be substantially reduced by preserving only a selected range of coefficients and ignoring those less important. The feature vector prepared in this way can constitute the input for machine learning algorithms or the implemented classifier. Such reduction in computational complexity opens the way for using solutions based on contour detection in real time.

An important aspect which requires attention are the distance dependencies between the *Points of Interest* (POI) which in an obvious way differentiate particular groups of visemes [28]. A sample illustration is shown in Fig. 16. Lips may be closed, spread, rounded or open and the visibility of the teeth may vary. Such diversity makes it possible to measure the distances between particular points located on the lips. The selection of the most representative distances is analyzed in further sections of the present paper. From the linguistic point of view the distances between the lip corners and between the highest point on the upper lip and the lowest point on the lower lip are important [27].

An appropriate description of this difference enables the selection of the complete set of parameters which are used in the recognition process.

In Figs. 17 and 18 three most important parameters used in many implementations during parameter extraction from the lip area are shown [1, 22, 28, 40]. They are geometrical parameters: the outer horizontal aperture, the outer vertical aperture and the angle of lip opening. Often the surface area inside the lip contour is also added. It should be emphasized that the parameters w and h must be normalized in order to make them independent of the individual features of a particular speaker and the location of the camera [13]. For such normalization the distance between the nose and the chin is often used. Another important parameter may also turn out to be the w/h ratio. An analogical analysis may be conducted

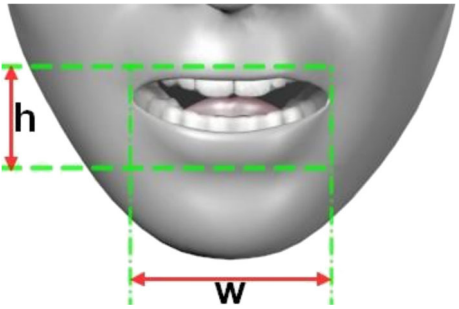


Fig. 17 Lip high and width marking

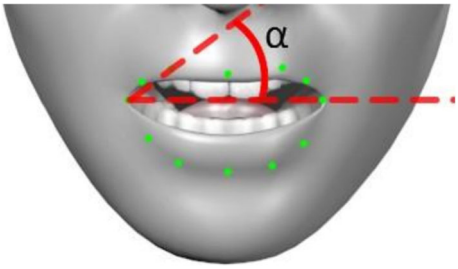


Fig. 18 Sample angle between the upper lip and the longest line segment on the horizontal axis

for the inner lip contour. In this way a relatively small set of parameters which will enable the training of the model may be obtained.

Other parameters may include the distances between the established center of gravity for the area inside the outer lip contour and the key points based on this contour. Such distances provide a lot of information about the lip opening and lip protrusion. The distances may then be added to the parameter vector [44].

It may also be interesting from the point of view of viseme description to consider the surface area of the teeth visible in a frame, which displays greater brightness than the adjacent elements in the oral cavity [16].

Another approach may involve encircling the area of detected lips in an ellipse. In order to place the area in an ellipse the points on the lip contour are used. The block circumscribed in this way is insensitive to location changes, e.g. rotation or a change in the size (visibility) of the upper and lower lip. The process of filtration and circumscription of lips on the ellipse can be represented in the following stages, as described in [19]. A sample result of implementing the above algorithm is presented in Fig. 19 which shows a visualization of an

Fig. 19 Sample result of algorithm application circumscribing an ellipse on the lip contour

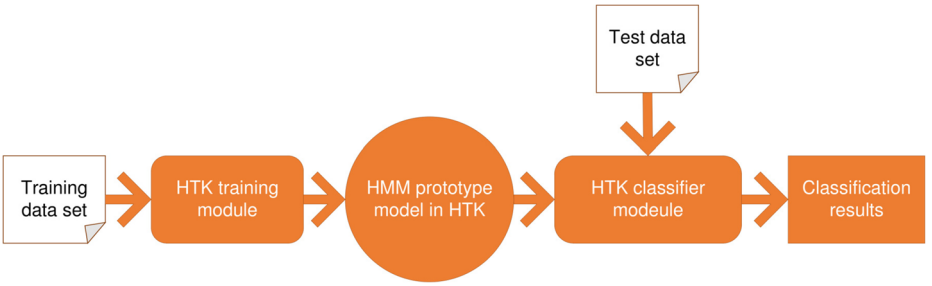


Fig. 20 Data flow while using the HTK package

ellipse circumscribed on the lips. The ellipse is marked together with the points on which it is circumscribed.

4 Testing environment

For the video recordings and data processing the *Python* programming language was used. It was also used for copying and data processing, frame extraction together with *ffmpeg* library and for calculating the geometrical and textual parameters.

Another module used during the research were appropriate codecs for carrying out the necessary operations. Thus, the library package *FFmpeg* (version 3.1.1 [36]) was downloaded and installed. Yet another library used in the analysis was the *OpenCV* package [37] downloaded and installed in version 2.4.12. This version has the most stable integration with the *Python* language. This library enables picture processing in real time as well as its effective scaling, trimming, filtering and calculating the parameters for particular frames. The library was used in histogram calculations and *Discrete Cosine Transform* (DCT) transform in the lip area. Am additional library was the *Numpy* package used for the processing of big sets of matrix data which shortened the time of analysis and improved the precision of calculations. The last library used in the analysis was the *Math* package which includes a vast set of mathematical operations.

Conducting the viseme recognition effectiveness tests requires the methods of machine learning. Two classifiers were used: the first one was based on Hidden Markov Models and the other on Support Vector Machine. Such approach makes it possible to check which of the classifiers is most effective and to compare the obtained results.

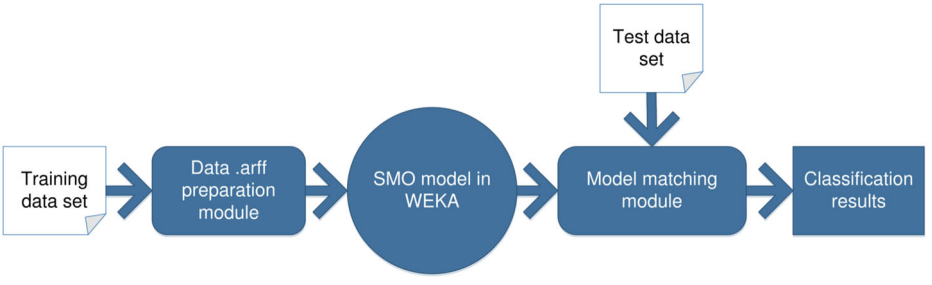


Fig. 21 Data flow while using the WEKA package

Table 2 Parameters of video files used in the recordings

Type	Picture
Duration	∼ 6 min
Size	∼ 3.5 GB
Codec	H264 – MPEG-4 AVC (part 10) (avc1)
Resolution	1088 × 1922
Picture resolution	1080 × 1920
Frames per sec.	100
Decoded format	Planar 4:4:4 YUV

The first classifier was the implementation of Hidden Markov Models in the HTK package (now version 3.4.1). Its schematic application is shown in Fig. 20.

Another classifier used for the analysis was the *Waikato Environment for Knowledge Analysis* package (WEKA) which implements a couple of algorithms of machine learning, large database processing and solves complex probability problems [38]. The package was written in JAVA. It was devised at the University of Waikato. The package of libraries is available on Open Source basis. In Fig. 21 the data flow in WEKA package is shown.

5 Data preparation and research procedure

It is explained in this section how the list of viseme groups was recorded and then broken down into the corresponding phonemes. Subsequently, the applied feature extraction techniques are summarized. The final task was to prepare a file containing all the parameters for each frame of the image used in the comparative analysis, employing two packages: the HTK package (HMM, Hidden Markov Models) and the WEKA package (SMO, Support Vector Machine), as is shown in the subsequent Section 6.

Fig. 22 Frames of recordings illustrating the realization of the /p/ viseme for different speakers **a**) Speaker21, **b**) Speaker22, **c**) Speaker23, **d**) Speaker26. Source: http://www.modality-corpus.org/

5.1 Material

The recordings included in a multimodal database for research on speech recognition and synthesis were used [36]. Four recordings of commands read by four different native speakers of English were selected for the present research on viseme recognition. The subjects were asked to adhere to their native Southern British English accent. The database included 230 carefully selected words of a potentially high degree of interaction with computer systems. The recordings together with the list of commands were used for further work on the extraction and parametrization of viseme frames; they were coded using *H*.264 *MPEG-4 AVC* codec. The complete set of picture parameters is shown in Table 2.

Figure 22 illustrates sample frames of the video recording showing the speakers producing the selected group of visemes. Speakers with a similar lip size were selected in order to compare them and minimize the error rate during the analysis. Each speaker had his own characteristic speaking expression, different speaking tempo and physiological conditionings. The visual data were also accompanied by complementary synchronous audio recordings. They were not necessary from the point of view of visual recordings; however, they facilitated the transcription of temporal dependencies between the beginning of the command and its end, which in turn enabled the extraction and analysis of particular visemes in the following stages of analysis.

The classification of visemes proposed in this study is based on articulatory similarities between certain phonemes. Particular groups and their features were presented in Table 3. It shows the articulatory label of a given group and an exemplary section of the labial ROI corresponding to this group. Each picture frame also illustrates the graphical prototypes of a particular group of visemes discussed in the theoretical part of the paper. A thorough analysis of Table 3 will enable the reader to grasp the differences between particular groups of visemes and facilitate the interpretation of the obtained results.

In order to carry out the extraction of static frames two scripts were prepared on the basis of the *FFMPEG* library. The aim of the first script was to change particular periods of time delimiting the duration of the uttered command into periods of time characteristic of a given phoneme. The application of the script were files which were formatted in a way which enabled the start of the other script responsible for uploading the video recording, the reading of the label files for particular phonemes and the phoneme $<->$ viseme mapping file. The script operation is based on establishing the duration of particular speech samples in a given command, uploading the temporal dependencies between the phonemes and finally eliciting the function which enables the extraction of the relevant static frames from the recordings.

The final result of the analyses was the obtainment of 440 frames. The number of the analyzed visemes amounts to eleven; thus each group contains 40 unique frames representing the visemes. Viseme-based recognition documented in 2016 by Heidenreich and Spratling brought 37.2 percent accuracy [11]. The parameters were calculated in result of feature extraction from the 3D-DCT representation. The examination and main conclusion were that the use of an extended training data set may not improve the score. The approach had a bad accuracy for some of viseme groups. At this stage the recorded frames were uploaded as input for the ROI detection algorithms of the speakers' face and the detection of the lip contour. The automatic detection of ROI in the context of viseme recognition was based on the *Active Appearance Model*. All picture frames at this stage were JPEG compressed with the highest quality coefficient 100%. The parameters of a sample frame are shown in Table 4.

The use of Intel RealSense scripts (based on the AAM method) enabled the obtainment of a file containing the location of points circumscribing the rectangle of the face and 20

Table 3 Selection, classification and characteristics of viseme groups

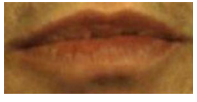
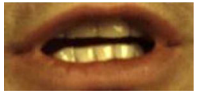
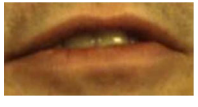
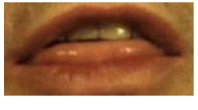
Class	Types of phonemes	Articulatory characteristics (place of articulation)	Sample realization (sample commands)	Sample frame of the labial ROI	Image description
W1	Consonantal: p, m, b	bilabial	*open, mute*		Lips are closed, possibly with a small aperture, slightly tense facial muscles, explosive character (short duration).
W2	Consonantal: t, d, n, s, z, l	alveolar	*as, ten*		Mouth is open, lips are full, teeth are visible and closed, long exposition.
W3	Consonantal: k, g, ŋ, ɬ	velar	*click, cut*		The upper lip is slightly constricted, teeth are visible with a small aperture between them, medium exposition.
W4	Consonantal: f, v	labiodental	*view, save*		The lips are constricted, directed upwards, in the shape of an upside-down letter *V*, potentially visible upper middle teeth.
W5	Consonantal: ʃ, ʒ, ʧ, ʤ	palato- alveolar	*check, flash*		The lips are lax, good visibility of the lower lip, possible 'cornet' shape, the tongue and the teeth are (potentially) visible.
W6	Consonantal: θ, ð	dental	*font, print*		The lips are open, upper teeth are visible through a wide aperture, lip corners are lax.

Table 3 (continued)

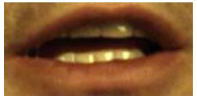
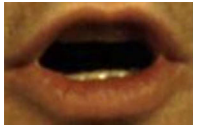
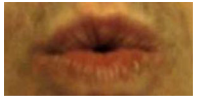
Class	Types of phonemes	Articulatory characteristics (place of articulation)	Sample realization (sample commands)	Sample frame of the labial ROI	Image description
W7	Vocalic: iː , , e, ɪ, eɪ, j	spread	*file, edit*		The lips are open and widely spread in the horizontal dimension, the teeth are visible.
W8	Vocalic: æ, a, ə	open-spread	*back,half*		The lips wide open and spread, the tongue is visible in the lower part of the mouth, possible visibility of the teeth.
W9	Vocalic: ɑː, a, ʌ, ə, h	open-neutral	*a.m., half*		The lips are open and apparently the biggest, the tongue and the upper teeth not visible, possible visibility of the lower teeth.
W10	Vocalic: uː, oː, ɒ, ʊ, ɔɪ	open-rounded	*one, up*		The lips are open, slightly flattened, poor visibility of the teeth, the tongue is invisible.
W11	Vocalic: ɜː, w, r, *kw*	protruding-rounded	*quarter, wife*		The lips are pressed together in the ‘cornet-like’ shape, the tongue and the teeth are invisible, possible lip protrusion in the ‘nozzle-like’ shape.

Table 4 Parameters of video files used in the recordings

Codec	JPEG image
Resolution	1080 × 1920
Horizontal pixel density	96dpi
Vertical piel density	96dpi
Number of bits per colour	24
Size	∼ 120 kB

points on the lip contour. The files were then ascribed to every picture representing a particular viseme. In order to optimize the efficiency of algorithms at this stage the graphic files representing the visemes were reduced by 50%. The resulting *.dat* file includes the location coordinates of particular points in the picture.

The file format contains the ROI and the resulting coordinates of the points. The first four digits (in bold type) describe the rectangle of the face: the distance from the left, the distance from the top, the width and the height. The following 40 digits are the pairs of x/y *coordinates* on the lip contour. The initial 12 points (underlined) are the coordinates of the outer lip, beginning from the left lip corner in the clockwise direction. The next 8 points (in italics, also beginning from the left lip corner) are the coordinates of the inner lip.

To visualize the results a script was used which drew the established points on particular frames. It was also possible to correct the location of certain points which had been determined incorrectly.

5.2 Feature extraction

The designation of coordinates of points on the lips contour mentioned in the previous subsection allowed in a subsequent step to calculate the geometric parameters. In turn, the designation of lips ROI allowed the calculation of the textural parameters. The calculation of the parameters for the data obtained in the earlier stages began by gathering them in one file. For this purpose, a script was developed whose aim was to identify and copy of the calculated points which defined the position of the speakers' lips along with the name of the frame to one created file. The feature calculation process implemented in the script is illustrated in Fig. 23. To calculate the textural parameters, frames were used in their original resolution instead of the reduced and compressed ones used for detecting the lip contour.

The first type of extracted parameters are geometric parameters. In order to calculate them the points describing the contour of the lips were used and the script which allowed the calculation of geometrical parameters, which can be divided into three types due to their origin: the distance, the angle and the surface. The principle of of the script operation is illustrated schematically in Fig. 24.

For each frame, 39 distance parameters were calculated. These parameters consist of the following:

- parameters representing the distance between the successive points on the outer periphery of the contour delineated on the speaker's lips relative to their sum, i.e. the circumference. 12 parameters were calculated for the outer contour,
- the same parameters calculated for the inner contour. 8 parameters were calculated for the internal contour,
- the distances of the straight lines connecting vertically the outer and inner contour points on the mouth in relation to the longest straight line in the horizontal plane. They

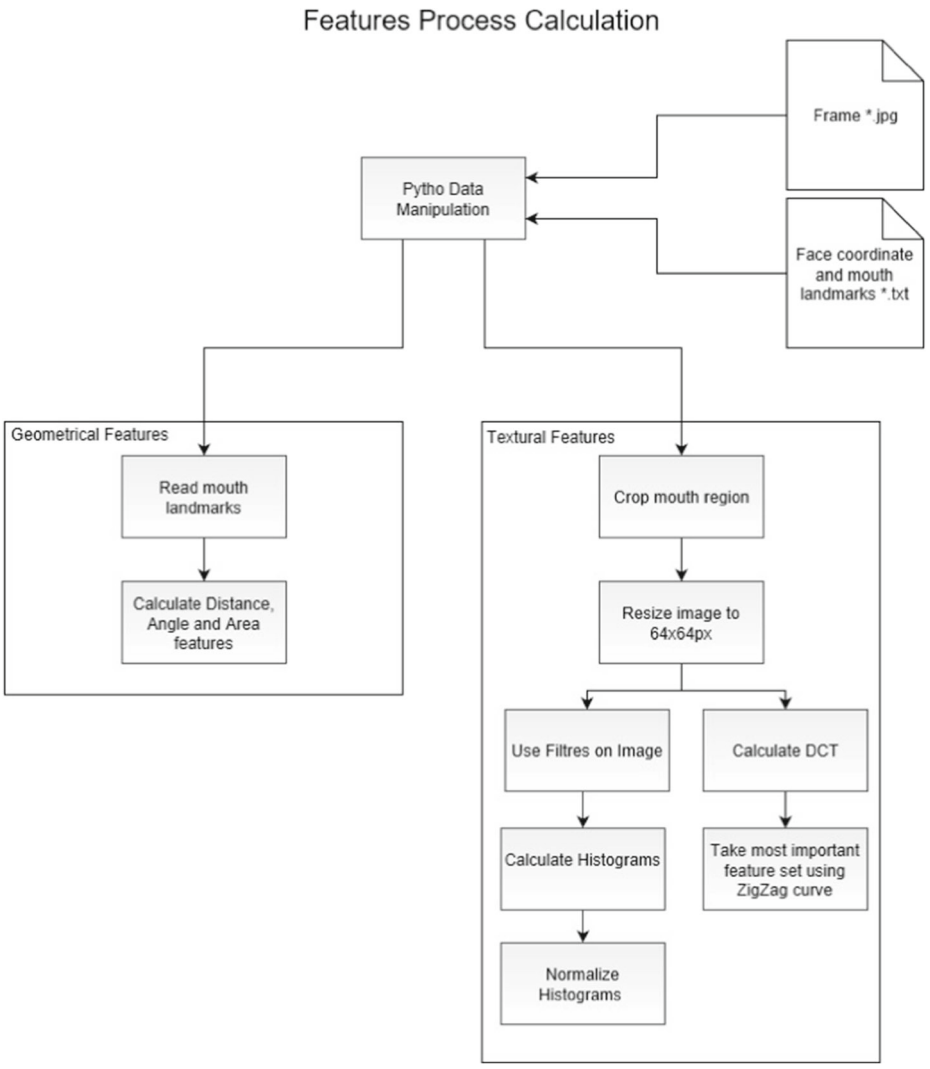


Fig. 23 Ilustration of feature calculation process

depict the maximum opening of the mouth in successive sections along the mouth, from left to right. The maximum opening found – 1 parameter. The opening for outer lips – 5 parameters. For inner lips – 3 parameters,

– the distances representing height versus maximum width, calculated for the exposure of the upper and lower lip while uttering a given viseme. They show the degree of lip exposure. 5 parameters were determined for the upper lip and 5 for the lower lip.

Moreover, 20 angle parameters were prepared. This type of parameters consists of the following values:

– 12 parameters calculated for the outer contour of the lips representing the values of the angles between successive points delineated on the lips in degrees. Two straight lines

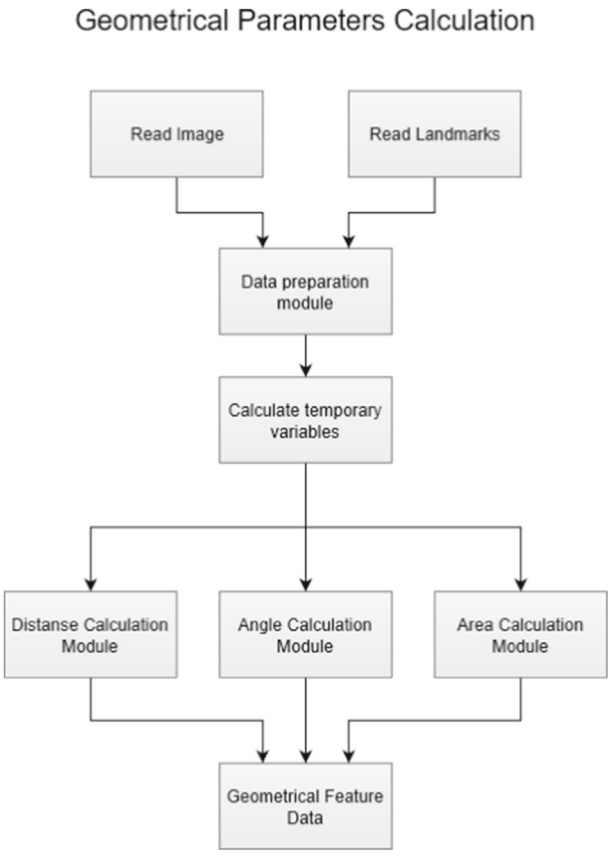


Fig. 24 Geometrical parameters calculation algorithm schema

were defined, drawn through successive points, which helped to calculate the angle values.
– 8 parameter values defined in a similar manner for the angles of the inner contour.

8 surface parameters were defined as well. These parameters represent the information about the visemes transferred in the areas of each frame image. These include the following calculated surface areas:

– the first parameter is the ratio of the area limited by the inner contour of the lips to the total area of the mouth, calculated for the outer contour,
– another element of the parameter vector is the ratio of the upper lip and lower lip area to the total area of the mouth,
– the next value defined is the ratio of the area limited by the inner lip contour to the surface of the upper lip
– similarly to the previous parameter, the following one is the ratio of the inner area to the lower lip area,
– another parameter is the ratio of the upper lip area to the lower lip,
– the next parameter is the ratio of the surface of the inner contour of the lips to the total surface of the lips,

– the last two parameters are the total area of the upper lip and the area inside the mouth to the surface of the lower lip and the sum of the lower lip and the area inside the mouth to the surface of the upper lip.

Textural parameters are the second type of parameters. They are based on the determination of histograms for ROI (*English for Region of Interest*) and of the DCT transform for subsequent frames of images. Textural parameters consist of the following types:

– 32 parameters representing the mouth histogram in shades of grayscale. An example of an area for the calculation of parameters is presented in Fig. 26a.
– 32 parameters that represent the mouth histogram within the HSV colour scale. The examples of *ROI* are presented in Fig. 26b.
– 32 parameters for the mouth image histogram in grayscale after applying the equalization. A sample image was shown in Fig. 26c.
– 32 parameters for the mouth image histogram in grayscale after processing via the Contrast Adaptive Histogram Equalization (CLAHE) [33]. *ROI* indexing a frame after filtering is illustrated in Fig. 26d.
– 32 parameters that represent the most significant values of DCT for the mouth area read in accordance with the Zig-Zag curve. A sample graph for the transform is presented in Fig. 26e.

The block diagram of the algorithm used for calculating textural parameters is presented in Fig. 25. Hassanat proposed and built an identification system based on the visualizing of the mouth. His research results show that the speaker authentication based on mouth movements can gain the security in the biometric systems [10]. The parameters prepared during the work presented in our paper can be also used in this kind of systems.

Sample results of contour detection and labial ROI can be observed in Fig. 26.

When optimizing the parameters obtained, a decision was made to trim the *ROI* of the mouth in the horizontal plane by about 10%. The objective here was to reduce the influence of pixels located in the corners of the analysed area. Then the coordinates of the rectangle depicting the relevant fragment of the mouth area were normalized to the constant adopted resolution of 64×64 pixels. The textural parameters were calculated for such reduced frame fragments.

160 textural parameters were defined. The histograms were carefully chosen in order to receive various values in the histograms obtained. They convey information about the number of pixels in the successive ranges of brightness. This enables to determine, inter alia, the exposure of the teeth, the tongue and the lips in an image frame.

The final task was to prepare a file containing all the parameters (a total of 227) for each frame of the image. The file pattern contains a label, the name of the parameters, the parameters of a given category, and the parameter values. It is presented in Table 5.

A different approach to the lipreading system operating on a word-level was proposed by Stafylakis et al. [31]. They prepare a deep learning neural network using approximately 2M parameters for each clip. The improved approach called VGG-M method allows for reaching a better score (6.8 percent higher) in word recognition compared to the previous state-of-the-art results. One of the conclusions was that the viseme-level systems allow an improvement of recognition of the start and the end part of the word, so the accuracy in the shortest words can be increased [31].

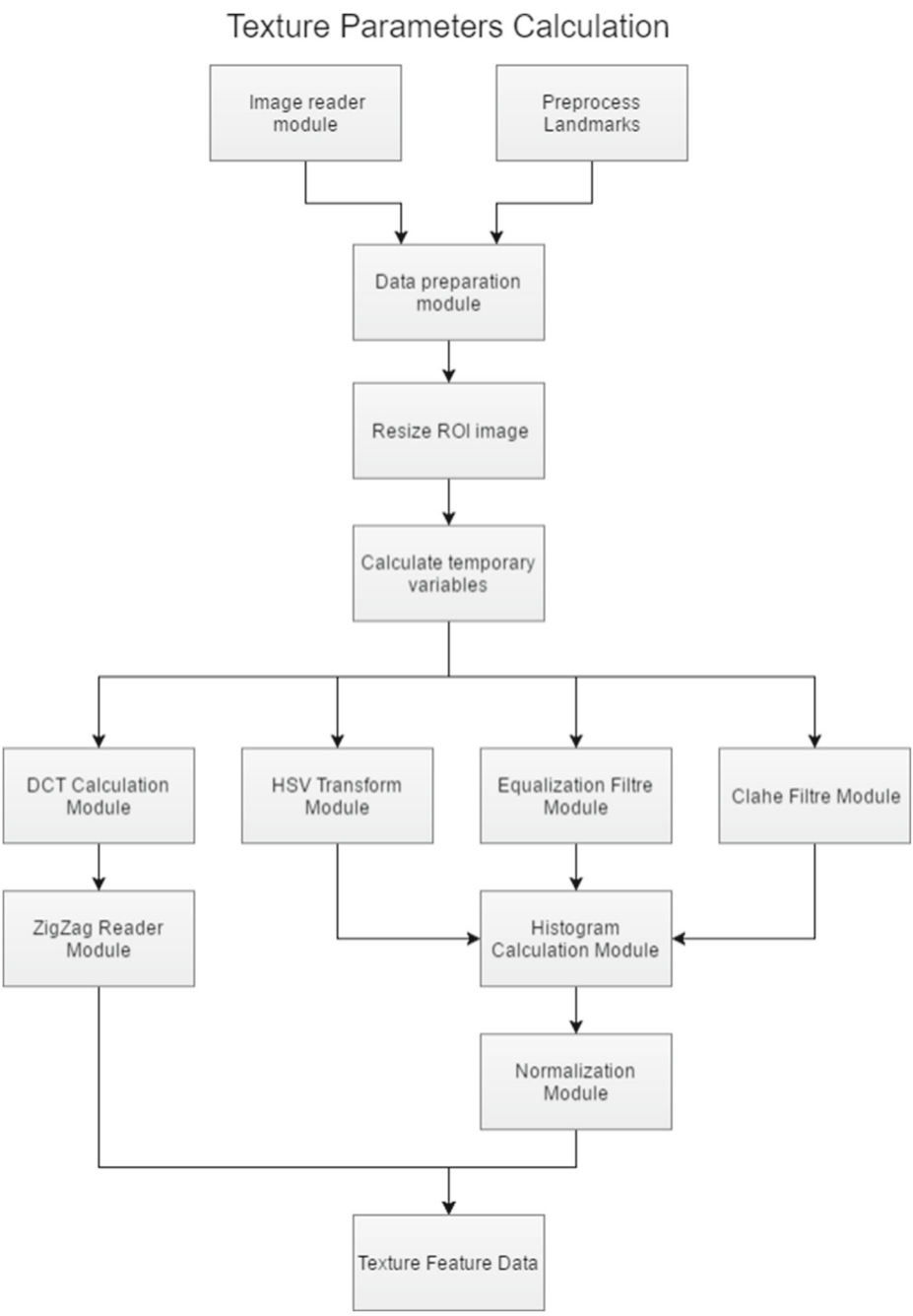


Fig. 25 Texture parameters calculation algorithm schema

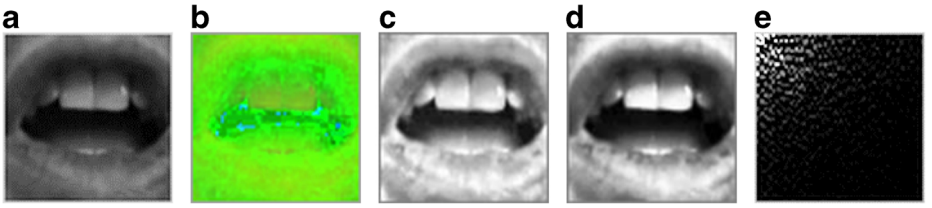


Fig. 26 Visualization of ROI frame analysis for the following images **a**) original, **b**) in HSV, **c**) after equalization, **d**) after filtering with CLAHE algorithm, **e**) DCT parameters

6 Experimental research

The calculated vector parameters for frames depicting a given viseme were divided into the training and the testing sets in line with the designed test scenarios. The pattern of action is shown in Fig. 27. The data was uploaded to a classifier by means of scripts. For this purpose, two classifiers were used:

- HTK package (HMM – Hidden Markov Models);
- WEKA package (SMO – Support Vector Machine extended by Sequential Minimal Optimization).

All tests were performed using a cross-validation mechanism. The mouth parameters were analysed in line with the aim of the study to determine the possibility of distinguishing individual speech elements – the visemes. A block diagram illustrating the choice of parameters is presented in Fig. 28. It assumes a check of detection efficiency for two classifiers, depending on the type of the parameters used. It was assumed that three test scenarios will be analyzed, making it possible to test the recognition effectiveness of a viseme class depending on the parameters. The data was properly prepared according to the structure of the files accepted as input for a given recognition system. For WEKA these are files with the extension *.arff*, while for HTK the files have the extension *.params*.

The parameters used in the HTK tool were prepared using the *VoiceBox* plug in MATLAB. The three scenarios tested were as follows:

- Scenario I: single parameter types (initial assessment of parameters carried out only for the SMO classifier);
- Scenario II: only the distance or textural parameters (for SMO and HMM);
- Scenario III: the use of the most effective set of parameters.

The analysis was carried out on the impact of the parameter type on the classification effectiveness of a given viseme class. The results will be discussed and conclusions from

Table 5 A sample of a file line containing features

Label	Dist-Params	Ang-Params	Area-Params	Hist-Grey	Hist-HSV	Hist-EQU	Hist-Clathe	DCT
	39	20	8	32	32	32	32	32
12_SPEAKER22_CONTROL_p_1_822919994.dat	0.0937	156.562	0.0426	1.1707	2.9536	117.162	0.0	0.065

Fig. 27 Division of frames per test scenarios

the results obtained will be presented. The WEKA package provided effectiveness metrics directly in percentages, while the HTK package provided result files containing individual projects of adjusting models to the test data. A script was developed and used to calculate the metrics.

Koller et al. presented a framework for the speaker-independent recognition of visemes to support the deaf people with their sign language communication [17]. They achieved 47.1 percent precision rate in the recognition attempts based on a dataset containing 180000 frames. Their research included the approach to the recognition of sequence of visemes. The conclusion of their work is that adding a dedicated viseme interpreting module to sign language recognition systems may gain their accuracy [21].

6.1 The first scenario (SMO)

The aim of the first scenario was to illustrate the extent to which the various types of parameters can be effective in the detection of the viseme class. SMO classifier training sessions were conducted for four speakers with the use of single parameter classes. The study allowed to draw conclusions about the advisability of the use of the analyzed parameter during the recognition of the viseme class as well as about their potential impact on their use in a mixed parameter class. The graphs in Fig. 29a-c show the efficacy results obtained for the SMO classifier from the WEKA package. At this stage, it was decided not to use the HTK package due to the limitations of the classifier, because it requires a more comprehensive data vector to create valid models for each of the classified viseme groups.

As is apparent from Fig. 29a, the distance parameters obtained showed the highest recall and precision for viseme classes W1, W4 and W11 and the lowest for W3, W6 and W9. This is due to the fact that the distance relations show the best results when the speaker utters the phoneme in which the mouth is arranged with lip closure. In turn, it poorly characterizes rounded wide open mouth. In addition, the method is very sensitive to the place of

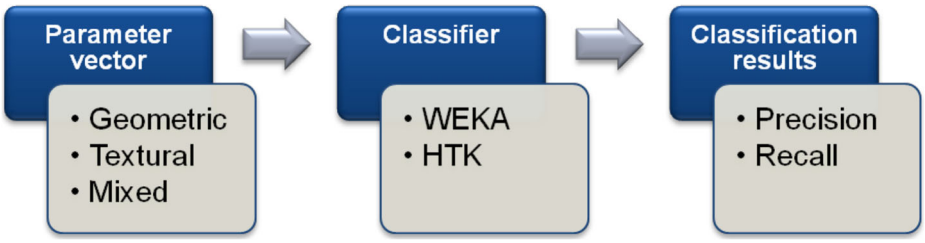


Fig. 28 The method of applying parameter vectors

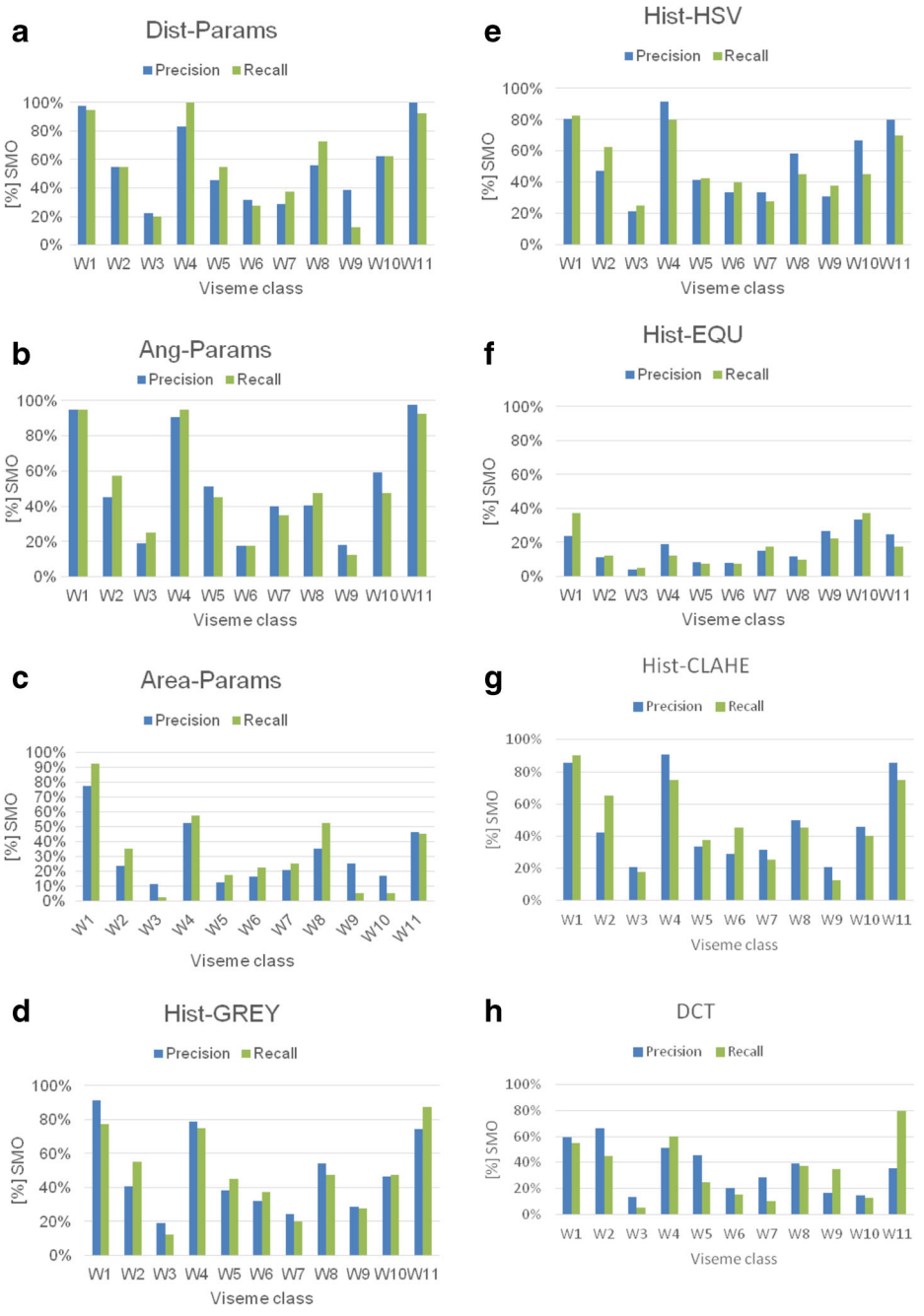


Fig. 29 Graphs showing the results of the first test scenario with the use of: **a**) distance parameters, **b**) angle parameters, **c**) surface parameters, **d**) parameters of the original ROI histogram, e) histogram parameters for HSV, **f**) parameters of the histogram after equalization, **g**) histogram parameters after LAHE filtering, **h**) DCT parameters

articulation, as it does not convey the information about the events occurring inside the lips (exposure of the teeth and tongue).

Figure 29b shows the results for the angle parameters. They show similar characteristics to the parameters of distance, because the return is at a similar level. However, the precision achieved is lower. The results for viseme classes where the frame shows the teeth were lower than in the previous one. The angle parameters have a low efficiency when recognizing classes where the lips wide open and rounded.

Figure 29c presents the results for the parameters indicating the area surface. They showed low efficacy in the detection of the viseme class. The exceptions include the W1 group (good efficiency of $\sim$ 70%) and W4 and W11 (average efficiency of $\sim$ 50%). They received a very low efficiency for W3, W9 and W10. The area parameters coped the least efficiently with the characteristics of the groups showing teeth within the image frame. They did not show efficacy because they are characterized by high sensitivity to the different physiognomy of the mouth area of the speakers (two speakers with a small mouth, one with a medium mouth and one with a large mouth).

The first tested textural parameter was a histogram of the original grayscale image. The recognition results are shown in Fig. 29d. It allowed to obtain a high precision in classes W1, W4 and W11. It proved to be effective in the detection of a large number of components with a similar dark shade (a large number of pixels of similar brightness saturation observed for the closed mouth as shown in an image). It poorly handled classes W3 and W7, where the teeth exposition plays an important role. Its effectiveness is low during the classification of the brighter shades. For other groups, the parameters proved to be effective at the level of $\sim$ 40%.

After transforming the original image to the HSV color scale and after calculating the histogram for the brightness component, the following results were obtained (Fig. 29e) which are characterized by detection rates higher by several percentage points for 9 viseme classes

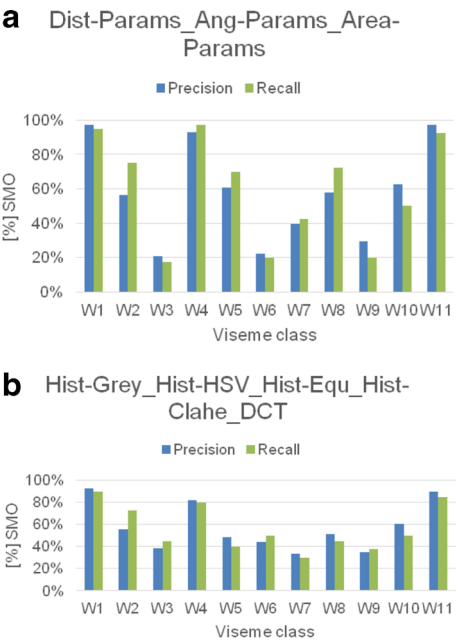


Fig. 30 Graphs showing the results for the set of geometric parameters (**a**) and for the set of textural parameters (**b**)

as compared to the results obtained for the original image histogram. This is due to a better representation of the value of brightness, which is presented directly, than of grayscale images. These parameters were to characterize the presence of individual elements, such as the tongue or the teeth in the analyzed mouth area of the speaker.

Testing the effectiveness of the viseme group classification using the parameters representing the values of the histogram for the image of the speaker's mouth after the equalization showed the ineffectiveness of this type of parameters. The results were the weakest among all the parameter types used. This is due to a weak correlation of image parameter values after equalization to the actual information of the unit of speech transferred. This transformation makes the histogram values stretch to the full range of the scale and in a way presents them as average values. This causes problems during the operation of the classifier in order to create models for each class. The chart showing the results for the parameters of the histogram after the equalization is shown in Fig. 29f.

The histogram values used, computed for the frame after filtering by CLAHE method as parameters, showed a good efficiency. The results were presented in Fig. 29g. The high efficiency for classes W1, W4 and W11 stems from the good separation of the parameters extracted for the dark areas in this histogram. These parameters, however, cope poorly with the presence of the teeth in the frame and wide-open mouth presented in ROI. The classifier obtained the weakest effectiveness precisely for these classes where the teeth and the tongue in the ROI area were visible.

The results obtained by calculating the content of the frequency components in the image (Fig. 29h) showed an average performance. Reducing the length of the vector to the 32 most significant components resulted in the loss of information about the high-frequency components that transfer data on the presence of the teeth in the frame and of widely open mouth. It would be moreover necessary to test the use of a longer vector of these features, e.g. after data processing via the PCA (*Principal Component Analysis*) method [21].

6.2 Presentation of results for the second scenario (SMO and HMM)

The second scenario followed the testing of the first scenario. The second scenario was designed to test the effectiveness of the combination of all the above parameters calculated, divided into two sets, taking into account class parameters. Therefore, two scenarios were

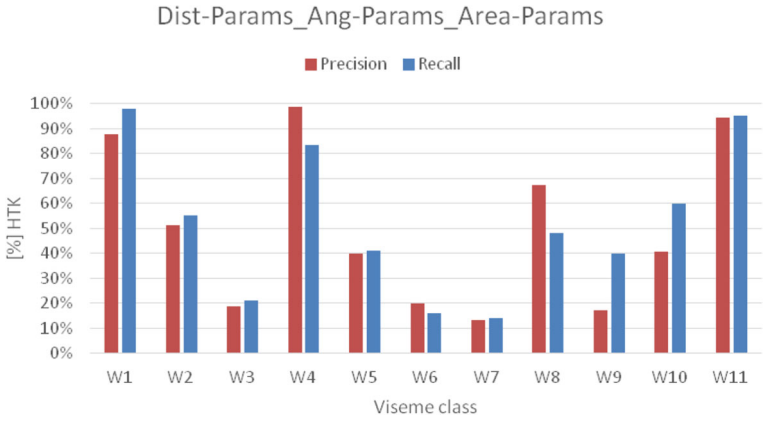


Fig. 31 Results for the set of geometric parameters

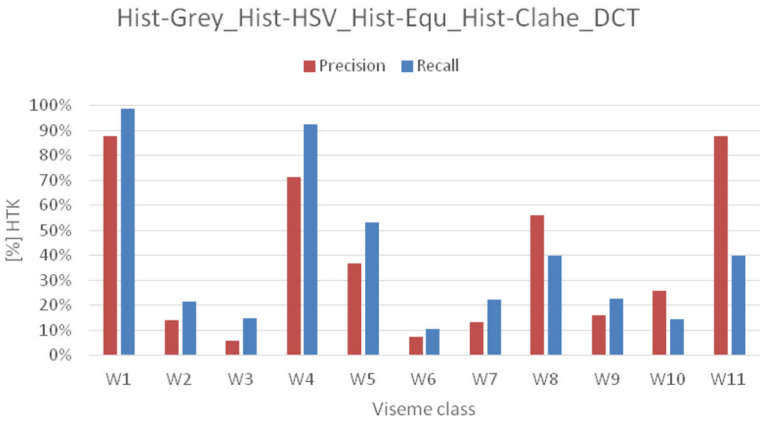


Fig. 32 Results for the set of textural parameters

tested; the first one for the geometrical parameters and the other one for the textural parameters. The results for the SMO classifier were presented in graphs in Fig. 30a and b. The results for HTK were presented in the diagrams in Figs. 31 and 32.

The use of the combination of geometric parameters yielded good results for some of the classes, including more than 90% efficiency for class W1, W4 and W11. This is a satisfactory result considering the amount of material used for training and tests. Furthermore, the average effectiveness rate of about 60% for classes W2, W5, W8 and W10 was obtained. It is important to note that the presented results were obtained for four different speakers. The parameters demonstrated a low efficacy in classes W3, W6 and W9. Classes W3 and W6 are somehow twin classes, where the difference is the place of articulation of the phoneme (not evident externally with the use of RGB cameras). The observation of the error matrices allows to conclude that the classifier had a problem distinguishing between these classes. However, it was wrong within their limits, so if these classes were considered as one, the obtained result would be 40% in terms of precision and recall.

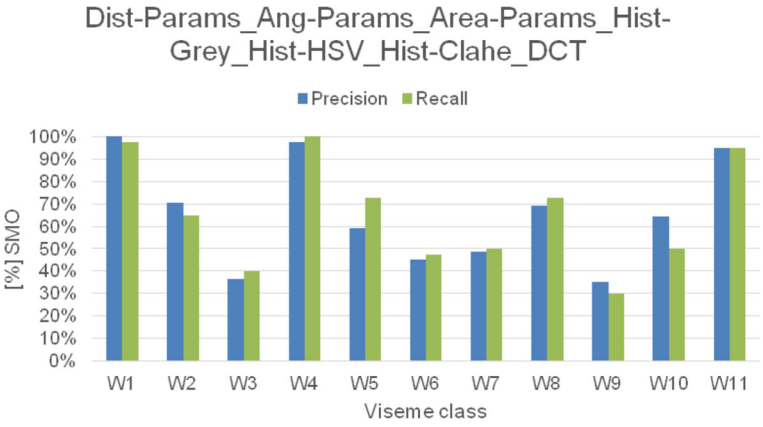


Fig. 33 SMO results for the most effective parameters

Using the group of textural parameters, good results were obtained in most of the classes. They demonstrated better efficacy in classes where the geometric parameters showed the lowest results. The standardization of ROI to 64x64 pixels for each frame image and then the calculation of the parameters helped reduce the classifier sensitivity to the physiognomy of the speakers. They coped the least efficiently with class W7, whose specificity is the greatest horizontal span of the mouth in all the groups. The application of the transformation to the standard definition removed a substantial part of this characteristics.

After the use of the geometric parameters as a set of training and test tools for HTK, the results obtained are presented in the chart in Fig. 31. Satisfactory effectiveness was obtained for the following viseme groups: W1, W4, W8 and W11. The calculated measures of classification accuracy for groups W2, W5, and W10 represent the mean efficiency of about 45%. In contrast, the groups W3, W6, W7 and W9 demonstrated a low efficacy. The classifier using Hidden Markov Models was adequately prepared to recognize the parameter type designated as *USER*. The results may be a bit biased due to the small amount of test and training data fed as the classifier input. The implementation of HMM in the HTK package requires a comprehensive set of training and test examples of precisely defined time dependencies. This caused problems when creating a suitable prototype in order to obtain optimally trained models. The results, however, legitimize conclusions about the quality of the analyzed geometric parameters. They showed better performance than the results for the textural parameters.

The graph presented in Fig. 32 demonstrates the results obtained for the KMM classifier for the set of textural parameters. Satisfactory efficiency of over 70% of classification for the following viseme groups was obtained: W1, W4 and W11. Groups W5 and W8 were recognized with average efficiency. Other groups demonstrated a low efficiency. The results of the HMM viseme classes classification groups demonstrated good efficacy in the separation of viseme classes where the mouth assume a very similar shape for each utterance in this group (regardless of the speaker). In the groups where the teeth exposure analyzed in the image frame was the main carrier of information on viseme group affiliation, HTK demonstrated a low efficacy. The distinction between the groups where the mouth was open also posed problems.

6.3 Presentation of results for the third scenario (SMO and HMM)

The third test scenario assumed the use of a combination of all parameters: geometric and textural, which showed the highest classification efficiency in the studies described in Section 6.1. The set of parameters adopted is analyzed in this section.

The graph in Fig. 33 shows the results obtained for a set of both geometric and textural parameters. The parameters used included distance, angle and surface ones as well as histograms calculated for the original grayscale image, HSV, the Clahe transformation, and for the vector of the most significant DCT coefficients. They demonstrated the highest efficiency in the classes W1, W2, W4, W5, W8, W10, and W11, achieving more than 60% efficiency. The results for these classes were considered satisfactory. Bearing in mind that the classes W3 and W6 can be put together in one class and analyzing the error matrix one can infer that this class could also have satisfactory efficiency at about 60%. Class W9 once again showed the lowest efficiency. The viseme class W9 was not adequately classified by any of the parameters analyzed. The problem with the parameterization of this class is due to the nature of the phonemes included in its composition, which, depending on the adjacent phonemes and the expressiveness of the speaker, demonstrates a high dynamic range of visual realizations.

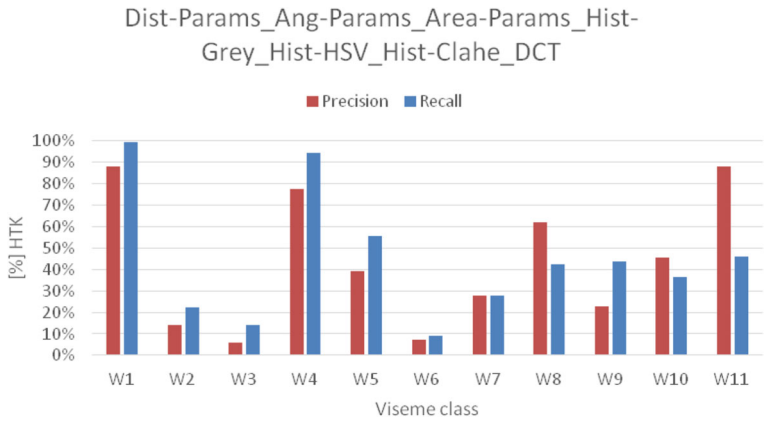


Fig. 34 HTK results for the most effective parameters

Figure 34 shows the results obtained for HMM using the most effective set of parameters. The viseme classes W1, W4, W8, and W11 showed a good efficiency. The results obtained for groups W5, W8 and W10 are at a medium level. Groups W2, W3 and W6 showed very low efficiency. In the case of group W2 a significant reduction in classification effectiveness was observed following the addition of textural parameters to the vector of geometric features. HMM cannot adequately fit the test data to models in the groups characterized by the presence of teeth in the analyzed ROI area. This may be due to insufficient data to establish an appropriate model. A more comprehensive training and test set should be used.

6.4 Summary of results for the scenarios and the classifiers

The overall efficiency for all tested sets of parameters for the SMO classifier from the WEKA package is presented in a single chart (Fig. 35). It may be observed that the analyzed different sets of parameters allow to achieve the same level of overall effectiveness. These sets provide similar performance characteristics for all the 11 groups of elements of speech, or visemes, analyzed.

In all the scenarios the best classification performance was obtained in the same viseme groups; by contrast, the worst results were obtained for the same viseme groups. However, the differences between the geometric and the textural parameters sometimes reached a few tens of percentage points. By optimizing the calculated parameters and adding the vectors of textural features only for the inner lip contour, one could obtain additional input data to create models characterized by a better separateness of the groups which currently produce the weakest results.

A summary of the results obtained for the HMM classifier from the HTK package is presented in Fig. 36. The average effectiveness for each scenario is at a similar level (about 50%). This classifier proves to be sensitive to a small amount of training data. The set of test frames should be bigger to explore possible changes in the results of viseme classification efficiency.

Conducting tests for individual types of parameters in the first scenario allowed an assessment of their impact on the detection of certain elements characteristic of each viseme group. The results obtained indicate that the parameters adequately describe the visemes of the groups W1, W4 and W11. The calculated textural parameters in conjunction with the

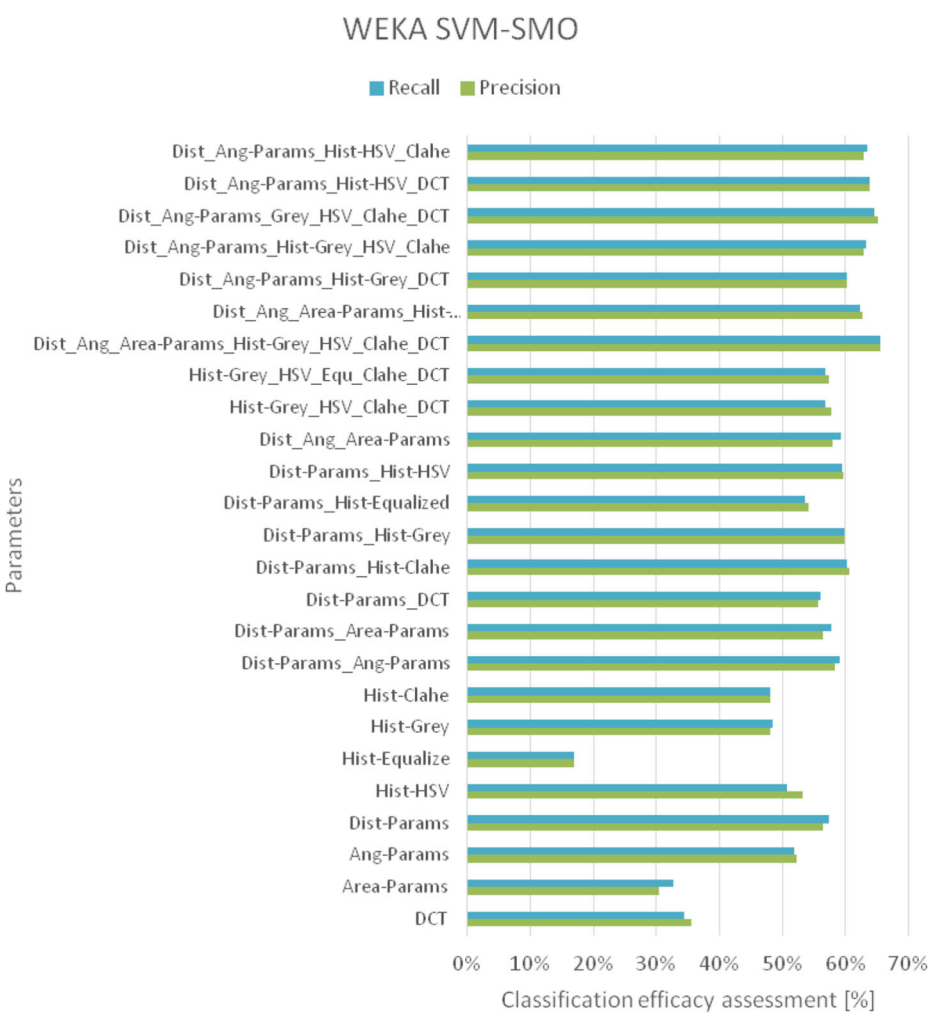


Fig. 35 Results of SMO classifier for the scenarios studied

geometric ones are able to cope well with groups W5 and W8. This indicates that they adequately reflect the presence of the tongue in the image frame. Of particular importance for the detection of the tongue are the histogram values for the image in the HSV scale. The parameters calculated for visemes from groups W2 and W7 have average performance due to the fact that they seem to be a little resistant to the appearance of a particular speaker's mouth when they are uttered. They are not able account for the small differences between these classes (e.g. width of the opening between the teeth) with sufficient accuracy. The phonemes included in these viseme groups show a high correlation with the adjacent speech fragments. Its nature is similar to the averaged image obtained as a result of calculating the average appearance of the lips in each viseme group. In these groups, the parameters poorly separate them from one another and from groups W3 and W6. The analysis of the error matrix of the results obtained by the classifiers legitimizes a conclusion that groups W3 and W6 are often erroneously classified within their boundaries. Group W9 is characterized by

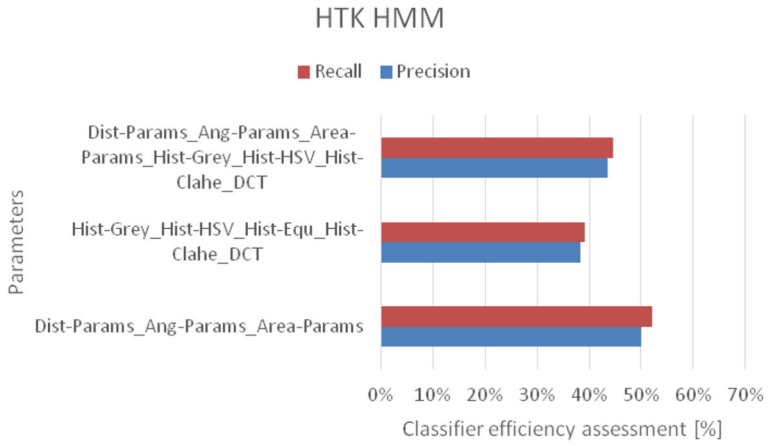


Fig. 36 Results of HMM classifier for the scenarios studied

high volatility in the way it is uttered by speakers, so it is hard to obtain satisfactory results using the parameters analyzed.

The results obtained for the SMO and HMM classifiers are similar in nature for each of the groups analyzed. The analyzed parameters allowed to obtain the best results for the SMO classifier. The selected set of features analyzed in section 6.3. achieved the highest effectiveness across all the tests carried out. The SMO can cope better with viseme separation within a test sample analyzed and is characterized by the lack of sensitivity to the size of the data set. The results obtained for the HMM were general the worse. The approach used during the tests assumed the use of a three-state prototype for models in the HTK core. Thus, it is possible that the models obtained are insufficiently accurate for the analyzed dataset. Successive model estimates did not differ too much from preceding ones as to probability values. The use of a prototype of a model with a higher number of states proved impossible. The HTK module calculating successive probabilities of transitions between the states of the model required the input of a more comprehensive set of training data. These problems were related to the configuration of the environment assuming the use of a model of input data of the *USER* type and top-down determination of the time relations between the frequency of the occurrence of the following labels (denoting a viseme) with the parameter vector correlated with it.

7 Conclusions and directions for further research

Although the algorithmic viseme (the smallest recognizable unit correlated with a particular realization of a given phoneme) recognition has been massively studied, there are no fully satisfactory results in the recognition of speech elements on the basis of lip picture analysis alone. A methodology was arrived at according to which phonemes are classified into the corresponding phoneme groups which are further assigned to appropriate classes of visemes. The different methods and approaches to this problem are then described in detail. Finally, a comparative analysis of their efficiency is performed. It was shown that the combination of geometrical and textural parameters enables a more efficient clustering in some of the newly defined groups.

A survey of viseme recognition methods was carried out and various ways of parameterization were examined. One of the tasks was also to compare the efficacy of selected algorithms of machine learning trained with parameters related to the mouth image. The influence of different types of parameters on the efficiency of recognition was extensively analyzed in the paper. Tests were organized according to three different scenarios:

- single parameter types (SMO) to illustrate to what extent the various types of parameters can be effective in the detection of a viseme class.
- distance-only (geometrical) or textural parameters (SMO and HMM) to test the effectiveness of the combination of all the parameters studied in the first experiment, divided into the two aforementioned groups (geometrical and textural)
- the use of the most effective set of parameters (SMO and HMM), assuming a combination of the previous parameters (geometric and textural).

So far few published works have examined feature vectors comparatively; therefore the results can serve as a basis for further analysis and for the development of an optimal way of extracting parameters from the area of the speaker's mouth. The suggested geometric parameters tend to model the viseme more generally as they were selected to reduce the influence of the shape/size of the speaker's mouth, while the parameters presented in the literature sometimes depend heavily on the speaker's individual physiognomic factors.

As it was stated above, one of the important results of the study was the preparation of the list of viseme groups, broken down into the corresponding phonemes. It was created as a result of the analysis of materials related to machine recognition and speech processing (in the context of the visual component) and the linguistic analysis of words belonging to the corpus used for multimodal recordings. The resulting division is different from the one most commonly used in the relevant literature, introducing a greater variety for vowel phonemes in the context of the classification adopted. Consequently, the viseme groups created can be used in other studies.

The main conclusion drawn from the analyzes is that the effective classification can be made for a given viseme. The study returned an average effectiveness of 65% for WEKA and 50% for HTK. The use of each of the classifiers allowed to obtain a similar mean classification efficiency within the viseme group for the parameter used. The calculated geometric and textural parameters and the use of both these types enabled a very efficient data clustering of 90% in viseme groups W1, W4 and W11. The prepared parameters also showed an efficacy of 65% for classes W2, W5, W8, and W10. The results obtained for groups W3, W6, W7 should be improved by fine-tuning of the parameter vector, more adequately carrying information about the location of the teeth in the analyzed frame. The poor efficiency of classification for group W9 is largely due to the variable manner of articulation of the sounds included in this group. The diversity of visual expressions requires the parameterization catering for a high dynamics of change in the appearance of the speakers' mouth, depending on the command uttered. This poses a challenge because of the huge impact of the unique characteristics of speakers' physiognomy for this class. A set of geometric parameters supplemented with textural parameters proved to be the most effective one; it can be further developed and optimized in order to improve the recognition efficiency. The directions for further research might involve the development of:

- the vector of distance parameters,
- the vector of angle parameters,
- the histogram calculated for HSV,
- the histogram after filtering by CLAHE,
- the parameters of the DCT transform.

Furthermore, the analysis could include the vectors of parameters obtained from the combination of the above ones, upon the use of PCA (*Principal Component Analysis*) algorithms. Reducing the vector dimension by using this algorithm could result in a better efficiency and assure the use of more parameters calculated for the DCT transform.

Additionally, one can also analyze the effectiveness of the parameters calculated for the averaged models created for each viseme group, e.g. through the use of algorithms of *Eigen-Face* type. The averaged models created in this way could be used to determine a new set of parameters. In order to better reflect the presence of the teeth one should obtain the textural parameters calculated for the inner contour of the lips. An interesting set of parameters could be the histograms of the entire surface of the mouth and, additionally, solely for the inner lip contour transformed to the shape of a quadrilateral (e.g. rectangle) by means of reverse parametrization. Reducing the impact of pixels that do not directly make up the mouth area could improve the results obtained for the textural parameters. This would allow, for example, to improve the exposure of the surface of the teeth (or the lack of thereof) in the image frame. The reflection of the teeth exposition could also be due to the geometric parameters calculated for the points whose coordinates should be determined on the contour of the teeth.

Bearing in mind the continuous nature of speech one should carry out tests on the effectiveness of the parameters for an increased number of frames fed into a classifier at fixed time intervals. This would enable an analysis of the results in the context of smooth transition between successive visemes, e.g. an analysis of three consecutive phonemes, or triphones, but in the context of their being mapped to visemes. This type of tests would facilitate the preparation of more accurate models for HTK for each of the viseme group.

One should also consider the possibility of extracting features from the interior of the speech apparatus (the movement and the position), e.g. three areas of the tongue inside the mouth, which are not visible in the RGB camera recordings. These features would allow the preparation of parameters that can improve the classification efficiency of viseme groups consisting of phonemes with a strong involvement of the tongue during the articulation of a given speech fragment. In this context, one could consider using data from a specialized device of electromagnetic articulography with its adequate parametrization or as shown in a recent paper by Yang et al. [43], one can also to employ emotional head motion predicting from prosodic and linguistic features or data acquired from a face motion capture device [24].

Nevertheless, the results obtained at this stage demonstrate that one can successfully carry out viseme classification using the SMO or HMM algorithms. The method of viseme division, along with a set of corresponding phonemes, and the methods for calculating the parameters allowed to indicate the directions in which to develop this field of expertise in order to arrive at highly efficient multimodal speech recognition systems.

Acknowledgements Research sponsored by the Polish National Science Centre, Dec. No. 2015/17/B/ST6/01874.

References

1. Alizadeh S, Boostani R, Asadpour V (2008) Lip feature extraction and reduction for HMMBased visual speech recognition system. Signal Processing ICSP 2008. 9th International Conference, Beijing

2. Cappelletta L, Harte N (2011) Viseme definitions comparison for visual-only speech recognition. European Signal Processing Conference, Barcelona
3. Cappelletta L, Harte N (2011) Phoneme-to-viseme mapping for visual speech recognition. 19th European Signal Processing Conference, Barcelona
4. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(6):681–685
5. Dalka P, Kostek B (2006) Vowel recognition based on acoustic and visual features. Arch Acoust 31(3):1–14
6. Dalka P, Bratoszewski P, Czyżewski A (2014) Visual Lip Contour Detection for the Purpose of Speech Recognition. In: International Conference of Signals and Electronic Systems (ICSES), Poznań
7. Dong L, Foo SW, Lian Y (2003) Modeling continuous visual speech using boosted viseme models. information, communications and signal processing, 2003 and fourth pacific rim conference on multimedia. In: Proceedings of the 2003 Joint Conference of the Fourth International Conference IEEE
8. Fernandez-Lopez A, Sukno FM (2017) Automatic viseme vocabulary construction to enhance continuous lip-reading. In: Proceedings 12th Intenrnational Conference on Computer Vision Theory and Applications, vol 5, Porto, pp 52–63
9. Jadczyk T, Ziolko M (2015) Audio-visual speech processing system for polish with dynamic bayesian network models. In: Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015) Barcelona, Spain, pp 13-14. Paper No. 343
10. Hassanat A (2014) Visual passwords using automatic lip reading. Int J Basic Appl Res (IJSBAR) 13:218–231
11. Heidenreich T, Spratling MW (2016) A three-dimensional approach to Visual Speech Recognition using Discrete Cosine Transforms, CoRR
12. Hojo H, Hamada N (2009) Mouth motion analysis with space-time interest points. In: TENCON 2009 – 2009 IEEE Region 10 Conference, Singapore
13. Kaynak MN, Zhi Q, Cheok AD, Sengupta K, Jian Z, Chi Chung K (2004) Analysis of lip geometric features for audio-visual speech recognition. In: IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans. IEEE
14. Kaucic R, Bynard D, Blake A (1996) Real-time lip trackers for use in audio-visual speech recognition. In: Integrated Audio-Visual Processing for Recognition, Synthesis and Communication, London
15. Kaucic R, Blake A (1998) Accurate, real-time, unadorned lip tracking, department of engineering science. Computer Vision, 1998. Sixth International Conference, Bombay
16. Krishnachandran M, Ayyappan S (2014) Investigation of effectiveness of ensemble features for visual lip reading. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi
17. Koller O, Ney H, Bowden R (2014) Read my lips: Continuous signer independent weakly supervised viseme recognition. In: Proceedings of ECCV 2014: 13th European Conference on Computer Vision, Zurich, pp 281–296. https://doi.org/10.1007/978-3-319-10590-1-19
18. Leszczynski M, Skarbek W (2005) Viseme recognition – a comparative study. In: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE
19. Li X, Kwan C (2005) Geometrical feature extraction for robust speech recognition. In: Signals, Systems and Computers, 2005. Conference Record of the Thirty-Ninth Asilomar Conference, Pacific Grove
20. Lucey P, Terrence M, Sridharan S (2004) Confusability of phonemes grouped according to their viseme classes in noisy environments. In: Proceedings of the 10th Australian International Conference on Speech Science & Technology, Sydney
21. Maeda S (2005) Face models based on a guided PCA of motion-capture data: Speaker dependent variability in /s/-/R/ contrast production. ZAS Pap Linguist 40:95–108
22. Mengjun W (2010) Geometrical and pixel based lip feature fusion in speech synthesis system driven by visual-speech. In: 2010 Second International Conference on Computational Intelligence and Natural Computing Proceedings (CINC), Wuhan
23. Multimodal AVSR corpus: http://www.modality-corpus.org/
24. McGowen V (2017) Facial Capture Lip-Sync. M. Sc. Thesis Rochester Institute of Technology
25. Ms Namrata D, Patel NM (2014) Phoneme and Viseme based Approach for Lip Synchronization. International Journal of Signal Processing, Image Processing and Pattern Recognition. SERSC
26. Neti C, Potamianos G, Luettin J, Matthews I, Glotin H, Vergyri D, Sison S, Mashari A, Zhou J (2000) Audio-visual speech recognition, Technical Report
27. Petajan E, Bischoff B, Bodoff D, Brooke M (1988) An improved automatic lipreading system to enhance speech recognition. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, pp 19–25
28. Sagheer A, Tsuruta N, Taniguchi R-I, Maeda S (2005) Visual speech features representation for automatic lip-reading. Acoustics, Speech, and Signal Processing

29. Sargın ME, Erzin E, Yemez Y, Tekalp AM (2005) Lip feature extraction based on audio-visual correlation. Signal Processing Conference, Antalya
30. Stegmann MB, Ersbřll BK, Larsen R (2003) FAME – A flexible appearance modelling environment. IEEE Trans Med Imaging 22(10):1319–133
31. Stafylakis T, Tzimiropoulos G (2017) Combining residual networks with LSTMs for lipreading, CoRR
32. Verbots tools Character Studio Visemes: verboots.com
33. Vyavahare AJ, Thool RC (2012) Segmentation using region growing algorithm based on CLAHE for medical images. In: IET Conference Proceedings Stevenage: The Institution of Engineering andamp; Technology
34. Wang X, Hao Y, Fu D, Yuan Ch (2008) ROI processing for visual features extraction in lip-reading. In: Conference Neural Networks & Signal Processing, Zhenjiang
35. Wang L, Wang X, Xu J (2010) Lip detection and tracking using variance based haar-like features and kalman filter. In: Fifth International Conference on Frontier of Computer Science and Technology, Changchun
36. Website of project Ffmpeg: http://ffmpeg.org (access date 15.04.2016)
37. Website of project Opencv: http://opencv.org (access date 20.04.2016)
38. Website of project Waikato Environment for Knowledge Analysis: http://www.cs.waikato.ac.nz/ml/weka (access date 10.05.2016)
39. WenJuan Y, YaLing L, MingHui D (2010) A real-time lip localization and tracking for lip reading. In: 3rd International Conference on Advanced Computer Theory and Engineering, Chengdu
40. Williams JJ, Rutledge JC, Garsteckit DC, Katsaggelos AK (1997) Frame rate and viseme analysis for multimedia applications. In: Multimedia Signal Processing. IEEE Workshop, Princeton
41. Wikipedia.org/wiki/viseme, date 03.01.2015
42. Xu M, Hu R (2006) Mouth shape sequence recognition based on speech phoneme recognition. In: Communications and Networking in China. ChinaCom first International Conference, Beijing
43. Yang M, Jiang J, Tao J, Mu K, Li H (2016) Emotional head motion predicting from prosodic and linguistic features. Multimed Tools Appl 75:5125–5146. https://doi.org/10.1007/s11042-016-3405-3
44. Zhang X, Mersereau RM, Clements M, Brown CC (2002) Visual speech feature extraction for improved speech recognition. In: 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Orlando

Dawid Jachimski, M. Sc., Eng., graduated as an engineer from Gdansk University of Technology, Faculty of Electronics, Telecommunication and Informatics, in the specialty of Multimedia Systems in 2015 and then was awarded his M.Sc. in the specialty Software Engineering and Databases in 2016. His first graduate work concerned the "Evaluation of practical application of audiovisual speech recognition" and for the other diploma (M. Sc. level) the subject was the "Examination of viseme recognition algorithms and visual lip features". He currently works for a company developing high accuracy synchronization systems in various network environments. His main skills also include complex system design, Python programming and data processing, analysis and visualisation. His research interests concern automatic speech recognition, synchronization and audio signal processing.

Andrzej Czyzewski, Ph. D., D. Sc., Eng. is a full professor at the Faculty of Electronics, Telecommunication and Informatics of Gdansk University of Technology. He is an author or a co-author of more than 600 scientific papers in international journals and conference proceedings. He has supervised more than 30 R&D projects funded by the Polish Government and participated in 7 European projects. He is also an author of 15 Polish and 7 international patents. He has extensive experience in soft computing algorithms and their applications in sound and image processing. He is a recipient of many prestigious awards, including a twotime First Prize of the Prime Minister of Poland for research achievements (in 2000 and in 2015). Andrzej Czyzewski chairs the Multimedia Systems Department in Gdansk University of Technology.

Tomasz Ciszewski, Ph. D., Associate Professor, works for Gdansk University of Technology, Faculty of Electronics, Telecommunication and Informatics and for University of Gdansk, Faculty of Languages, Institute of English and American studies. He is a University of Łódź graduate (1995) and his PhD thesis (2000) was devoted to phonological analysis of the English stress system in a non-linear conditions-and-parameters? approach. He is an author of several papers published in domestic and international journals and conference proceedings on English phonetics and theoretical phonology. He also published two books: The English Stress System: Conditions and Parameters and The Anatomy of the English Metrical Foot: Acoustics, Perception and Structure (Peter Lang Publishing Group). In 2012 he was awarded the University of Cambridge Corbridge Trust Scholarship and the Ministry of Science and Higher Education research (2011). Tomasz Ciszewski is also the chair of the Interdisciplinary Laboratory for Speech Analysis and Speech Processing (University of Gdansk).