BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 60, No. 2, 2012
DOI: 10.2478/v10175-012-0041-6

Characteristics of the use of coupled hidden Markov models for audio-visual Polish speech recognition

M. KUBANEK*, J. BOBULSKI, and L. ADRJANOWICZ

Institute of Computer and Information Sciences, Czestochowa University of Technology, 73 Dąbrowskiego St., 42-200 Częstochowa, Poland

Abstract. This paper focuses on combining audio-visual signals for Polish speech recognition in conditions of the highly disturbed audio speech signal. Recognition of audio-visual speech was based on combined hidden Markov models (CHMM). The described methods were developed for a single isolated command, nevertheless their effectiveness indicated that they would also work similarly in continuous audio-visual speech recognition. The problem of a visual speech analysis is very difficult and computationally demanding, mostly because of an extreme amount of data that needs to be processed. Therefore, the method of audio-video speech recognition is used only while the audio-speech signal is exposed to a considerable level of distortion. There are proposed the authors' own methods of the lip edges detection and a visual characteristic extraction in this paper. Moreover, the method of fusing speech characteristics for an audio-video signal was proposed and tested. A significant increase of recognition effectiveness and processing speed were noted during tests – for properly selected CHMM parameters and an adequate codebook size, besides the use of the appropriate fusion of audio-visual characteristics. The experimental results were very promising and close to those achieved by leading scientists in the field of audio-visual speech recognition.

Key words: coupled hidden Markov models, audio-visual speech recognition, lip reading.

1. Introduction

There are used two separate streams of information in audio-visual speech recognition, each for every signal, in contrast to only one in audio speech. The combination of these streams should provide a better performance in comparison with modern solutions using each source separately. Although the use of visual features for a robust speech recognition system appears to be natural, there are several questions that need to be answered, such as: what a robust set of visual features is, what the best mean of audio and visual feature integration is, what represents the best model for audio-visual data.

In works [1–2] the authors presented an analysis of the efficient lip reading method for various languages. First, they applied an active appearance model, and simultaneously extracted the external and internal lip contours. Furthermore, teeth and an internal lip region were detected. Various features from five regions were fed to the recognition process. There were selected – as recognition targets - four languages with 20 words recorded for each of them. The proposed analysis of a feature trajectory based on three shapes (features, area with aspect ratio of internal lip region and area of intraoral region) provided the highest recognition rates of 93.6% in comparison with traditional methods and other regions [3]. In other works [4, 5] the authors presented the Bayesian model of optimal cue integration for the lip reading, where words were regarded as points in a multidimensional space and word recognition was a probabilistic inference process. While the dimensionality of the feature space was low the Bayesian model predicted an inverse effectiveness. On the other hand, while the dimensionality was high, the enhancement was maximal at intermediate auditory noise levels. Moreover, when the auditory and visual stimuli differed slightly in high noise, the model made a counterintuitive prediction: as sound quality increases, the proportion of reported words corresponding to the visual stimulus should first increase and then decrease.

In fact they confirmed this prediction in a behavioral experiment and concluded that auditory-visual speech perception obeyed the same notion of optimality previously observed only for simple multisensory stimuli [3]. The authors [6,7] proposed a real-time lip-reading method in smart phone environment, where resources were limited to existing PC environment. Therefore it was hard to achieve real-time lip-reading. In order to solve this problem they proposed the lip area detection method and feature extraction method suitable for smartphone environment. To find the accurate lip area the face area was detected by means of face colour information and eyes were located to detect the lip area with the geometrical relation. Then there were applied histogram matching, lip folding and RASTA filter - to extract the outstanding features of the lip area in terms of light changes according to the surrounding. Then extracted features were used during the recognition process. They showed that the changes were recognized almost in real-time, and 30 out of 50 words were recognized. That indicated about 60% recognition rates [3].

In works [8–9], the authors presented an integrated AVSR system, where noise tolerance was improved through enhancing the performance of three main components of the system. First, subsystem visual performance was improved by means of stochastic optimization methods for the hidden Markov models. Second, a new method of speech dynamic analysis was proposed, which improved acoustic efficiency. Third, an

*e-mail: mariusz.kubanek@icis.pcz.pl

efficient integration of both signal streams was used to determine final robust recognition results by utilizing neural networks.

This paper presents a study of an optimal selection of CHMM parameters settings with combination of codebook size and the use of this method for audio-visual Polish speech recognition. The described method was proposed to eliminate the negative influence of external factors on audio speech signal. The recognition was based only on isolated words, but the effectiveness of this method could be transferred to continuous speech recognition systems. The difficult speech video image analysis was used only in situations where the audio speech signals were exposed to disturbances. In systems operating in a quiet environment, one should not use the audio visual speech recognition, because it slows the whole system work. The mechanism of cepstral speech analysis was applied for extraction of person's speech audio features. The mechanism used a bank of amplitude-frequency filters with characteristics similar to human hearing. Besides that, twenty-dimensional MFCC (Mel Frequency Cepstral Coefficients) were used as the standard audio features for acoustic speech recognition. In addition own methods were created to determine the beginning and the end of isolated words in audio speech signal. Moreover, Lloyd algorithm for vector quantization was applied. Finally, automatic methods of face, eyes and region of mouth detection were used for visual feature extracting. The visual features were: the corners, outside edges of the lips, and the visible tongue.

2. Preprocessing of input audio signal

It is necessary – in systems of isolated words recognition – during recordings, to make a short-lived but clear pauses in form of silence among individual words. In case of this kind of recognition, after a preliminary filtration of a signal the next stage is to create a clean and proper audio signal, through removal silence at the beginning and at the end of a signal.

An identification, which frames are to be rejected, is not simple to determine whether energy matches the frame condition. There can appear instantaneous power spikes before the beginning of a useful signal. In most cases they are related to interference caused by the environment in which the signal recording takes place. Therefore, a more complex search is needed to determine the beginning and the end of the audio signal. For this reason the system uses two more parameters of the LRP (the initial number of frames) and LRK (the final number of frames). The first one specifies the number of frames consecutively, which energy satisfies the condition. If that number of frames is found, then the first of these frames is marked as the beginning of the audio signal. The second parameter specifies the number of frames consecutively, the power does not satisfy the condition. If that number of frames is found, then the first of these frames is determined as the end of the audio signal. If the required number of frames is not found before reaching the last frame, then the end of the signal is assigned to the frame, the first condition is not fulfilled. Assuming that the parameters of LRP and LRK will be set to the value 3, may be illustrated by examples from the energy distribution frames fulfilling this condition (see Fig. 1).

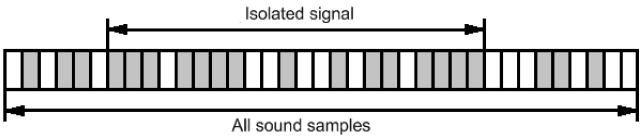

Fig. 1. The sequence or the way of searching LRP and LRK

Because the speech signal is not stationary, what results from dynamic properties of human speech, next stage depends on the use of division of entrance signal onto stationary frame boxes [10]. The signal is stationary in short temporary partitions (10 ± 30 ms) [11]. Every such a stationary frame box was replaced by the symbol of observation in the process of creation the observation vectors. In the created system it was accepted the length of every frame box equal 30 ms. To keep the signal stationary a method of delaying next frame boxes was applied. As a result every next frame box is sewing on previous with delay.

3. Encoding the signal using Cepstral analysis and Lloyd algorithm

Speech processing applications require specific representation of the speech information. The speech signal may be analysed in two main ways: as a signal generated from the speech signal based on voice characteristics [12], and registered as a recognized signal. A wide range of possibilities exists for parametrically representation of the speech signal. Among these the most important parametric representation of the speech is a short time spectral envelope [10, 13]. Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) spectral analysis models have been used widely for speech recognition applications. Usually together with MFCC coefficients, first and second order derivatives are also used to take into account the dynamic evolution of the speech signal, which carries relevant information for the speech recognition.

For the extraction of audio features, we use Mel Frequency Cepstral Coefficients (MFCC) to the analysis of the audio speech. In the process of speech signals perception the human ear makes non-linear (in frequency domain) spectrum analysis of this signal. Cepstral analysis using filter bank consists of passing the signal through a spectral band pass filter bank before switching into the field frequency. In order to adapt the characteristics of filters, frequency scale is converted into mel-scale using the following formula:

$$f_{mel} = 2595 \log_{10}(1 + f_{Hz}/700), \tag{1}$$

where f_{mel} – mel scale, f_{Hz} – frequency in the normal linear frequency scale.

Characteristics of filters copy the human auditory system. The filters had triangular bandpass frequency responses. The bands of filters were spaced linearly for bandwidth below 1000 Hz and increased logarithmically after the 1000 Hz. In the mel-frequency scaling, all the filter bands had the same

width, which were equal to the intended characteristics of the filters, when they were in normal frequency scaling.

Spectrum of signal of every frame boxes obtained by Fast Fourier Transform (FFT) comes under process of filtration by bank of filters. The next step is to calculate the members of each filter by multiplying the filter's amplitude with the average power spectrum of the corresponding frequency of the voice input. The summation of all members of filters is:

$$S_k = \sum_{n=0}^{(N/2)-1} (P_n \cdot A_{k,n}), \tag{2}$$

where S_k – power spectrum coefficients, N – the total number of samples in the framework, P_n – average power spectrum of the input sound, $A_{k,n}$ – amplitude.

The filter bank parameters are intricately linked, causing deterioration in recognition performance, even if one assumes the independence of the parameters in the vector of observations. The improvement of the quality of recognition may be achieved through the use of cepstral transform of filter bank parameters, involving the appointment of cepstral coefficients in mel scale MFCC as discrete cosine transformations of the logarithms of filter bank parameters according to the relationship:

$$MFCC_n = \sum_{k=1}^{K} (\log S_k) \cos\left[n(k-0.5)\frac{\pi}{K+1}\right], \quad \text{for} \quad n = 1...N, \tag{3}$$

where S_k – power spectrum coefficients, K – a required number of cepstrum coefficients, N – a number of filters in a filter bank.

An additional advantage of MFCC coefficients is decoupled speech signal from the influence of the transmission channel. Assuming that the transmission channel is a linear dynamic system, signal at its output is the convolution of the speech signal (input) and the impulse response of the system. Thus, the specter of a distorted speech signal (output) is the product of the speech signal spectrum and frequency characteristics of the transmission channel. The impact of the transmission channel is reduced by subtracting the average value of MFCC coefficients from all observation vectors in the field of cepstrum. In fact, the average value is estimated using a limited amount of data, and subtract the mean operation does not eliminate the influence of the transmission channel for the signal. Nevertheless, this simple technique is very effective in practice, to compensate for long-term effects on the spectrum, caused by different microphones and transmission channels used during the registration process of speech signals. To perform the cepstral normalization of the average value, it is an average value of each cepstrum coefficient for all speech learning.

The frequency band, which analyses the signal may cover the entire frequency range, or may be limited (for example, to reject the frequency ranges). In the latter case one should ask the lower and upper frequency bands analysed, the present number of channels in the filter bank to be spread evenly along the mel-scale range in this way, that the lower frequency bands will coincide with the lower cutoff of the first filter, and the upper frequency band is the upper cutoff of the last filter.

The use of loss compression, the data generated by the source must be represented by one of the small number of code words. The number of different data is generally much larger than the number used to represent those code words. Vector quantization is performed in the process of codebook based on a set of input records, covering the whole space for a given problem and the user. Vector quantization is to assign a suitable symbol of each frame of speech signal. In speech recognition systems based on HMM, each frame represented by a vector of observation is coded as a symbol of observation. There was applied Lloyd algorithm to vector quantization.

4. The audio-visual CHMM

A CHMM may be seen as a collection of hidden Markov models (HMM), one for each data stream, where the hidden backbone nodes at time t for each HMM are conditioned by the backbone nodes at time $t-1$ for all the related HMMs. The squares represent the hidden discrete nodes (backbone and mixture nodes) while the circles describe the continuous observable nodes. Unlike the independent HMM used for the audio-visual data, the CHMM can capture the interactions between the audio and video streams through the transition probabilities between the backbone nodes. In the system presented in this paper, the audio-visual CHMM allows for asynchrony in the audio and visual states but forces them to be synchronized at the model boundaries. In addition, with the coupled HMM, the likelihood of audio and video observation is computed independently, significantly reducing the parameter space and complexity of the model compared to the models that require the connection of the audio and visual observations [14].

The parameters of a CHMM are defined below:

$$\pi_0^c(i) = P(q_t^c = i), \tag{4}$$

$$b_t^c(i) = P(O_t^c | q_t^c = i,) \tag{5}$$

$$a_{i|j,k}^c = P\left(q_t^c = i | q_{t-1}^0 = j, q_{t-1}^1 = k\right), \tag{6}$$

where q_t^c is the state of the couple node in the c-th stream at time t.

In a continuous mixture with Gaussian components, the probabilities of the observed nodes are given by:

$$b_t^c(i) = \sum_{m=1}^{M_i^c} w_{i,m}^c N\left(O_t^c, \mu_{i,m}^c, U_{i,m}^c\right), \tag{7}$$

where $\mu_{i,m}^c$ and $U_{i,m}^c$ are the mean and covariance matrix of the i-th state of a coupled node, and m-th component of the associated mixture node in the c-th channel. M_i^c is the number of mixtures corresponding to the i-th state of a coupled node in the c-th stream and the weight $w_{i,m}^c$ represents the conditional probability:

$$P(s_t^c = m | q_t^c = i), \tag{8}$$

where s_t^c is the component of the mixture node in the c-th stream at time t. Unlike isolated word audio-visual speech recognition where one CHMM is used to model each audio-visual word, in audio-visual continuous speech recognition, each CHMM models one of the possible phoneme – visual pairs as defined in [15].

The training of the CHMM parameters for the system is performed in two stages and is an extension of the training used in audio-only speech recognition. In the first stage, the CHMM parameters are estimated for isolated phoneme-visual pairs. In this stage, the training sequences are labeled using an audio-only speech recognition system, and the phoneme-visual correspondence tables. The parameters of the isolated phoneme-visual CHMMs are estimated first using the Viterbi-based initialization followed by the estimation-maximization (EM) algorithm. To deal with the requirements of speech recognition systems, two additional CHMMs are trained to model the silence between consecutive words and sentences. In the second stage, the parameters of the CHMMs, estimated individually in the first stage, are refined through the embedded training of all CHMM from audio-visual speech. In this stage, the labels of the training sequences consist only of the sequence of phoneme-visual with all boundary information being ignored. Each of the models obtained in the first stage are extended with one entry and one exit non-emitting states-in a way similar to the embedded training for HMMs. The use of the non-emitting states also enforces the phoneme-visual synchrony at the model boundaries [14].

The embedded training follows the steps of the EM algorithm for continuous audio-visual speech, and is described by the following:

E step. The forward probability:

$$\alpha_t(i,j) = P\left(O_1,...,O_t, q_t^0 = i, q_t^1 = j\right) \tag{9}$$

and the backward probability:

$$\beta_t(i,j) = P\left(O_t+1,...,O_T | q_t^0 = i, q_t^1 = j\right) \tag{10}$$

are computed. Starting with the initial conditions:

$$\alpha_1(i,j) = \pi_1^0(i)\,\pi_1^1(j)\,b_1^0(i)\,b_1^1(j) \tag{11}$$

the forward probabilities are computed recursively from:

$$\alpha_i(i,j) = b_{t-1}^0(i)\,b_{t-1}^1(j)\sum_{l,k} a_{i,j|l,k}\alpha_{t-1}(l,k) \tag{12}$$

for $t = 2,3,...,T$. Similarly, from the initial conditions:

$$\beta_T(i,j) = 1 \tag{13}$$

the backward probabilities are computed recursively from:

$$\beta_t(i,j) = \sum_{l,k} b_{t+1}^0(l)\,b_{t+1}^1(k)\,a_{l,k|i,j}\beta_{t-1}(l,k) \tag{14}$$

for $t = T-1, T-2,...,1$ where i,j are the states of the audio and video chain respectively and $a_{i,j|k,l}$ = is the transition probabilities between the set of audio-visual states i,j and k,l. The probability of the r-th observation sequence:

$$O^r = [O_1^r,...,O_{T\,r}^r] \tag{15}$$

is computed as:

$$P_r = \alpha_{T\,r}(N,M) = \beta_1(1,1), \tag{16}$$

where N,M are the number of states in the audio and video chain respectively and T_r is the length of the observation sequence O_r.

M step. The forward and backward probabilities obtained in the E step are used to re-estimate the state parameters as follows:

$$\widetilde{\mu}_{i,m}^c = \frac{\Sigma_r\Sigma_t\gamma_t^{r,c}(i,m)\,O_t^r}{\Sigma_r\Sigma_t\gamma_t^{r,c}(i,m)}, \tag{17}$$

$$\widetilde{U}_{i,m}^c = \frac{\Sigma_r\Sigma_t\gamma_t^{r,c}(i,m)\left(O_t^r-\mu_{i,m}^c\right)\left(O_t^r-\mu_{i,m}^c\right)'}{\Sigma_r\Sigma_t\gamma_t^{r,c}(i,m)}, \tag{18}$$

$$\widetilde{w}_{i,m}^c = \frac{\Sigma_r\Sigma_t\gamma_t^{r,c}(i,m)}{\Sigma_r\Sigma_t\Sigma_m\gamma_t^{r,c}(i,m)}, \tag{19}$$

where

$$\gamma_t^{r,c}(i,m) = \frac{\Sigma_j\frac{1}{P_r}\alpha_t^r(i,j)\,\beta_t^r(i,j)\,w_{i,m}^c N\left(O_t^r,\mu_{i,m}^c,U_{i,m}^c\right)}{\Sigma_{i,j}\frac{1}{P_r}\alpha_t(i,j)\,\beta_t(i,j)\,\Sigma_m w_{i,m}^c N\left(O_t^r,\mu_{i,m}^c,U_{i,m}^c\right)}. \tag{20}$$

The state transition probabilities may be estimated using:

$$\widetilde{a}_{i|k,l}^{0,1} = \frac{\Sigma_r\frac{1}{P_r}\Sigma_t\alpha_t^r(k,l)\,a_{i|k,l}b_t^{0,1}(i)\,\Sigma_j\beta_{t+1}^r(i,j)\,b_t^{1,0}(j)}{\Sigma_r\frac{1}{P_r}\Sigma_t\alpha_t^r(k,l)\,\beta_t^r(k,l)}. \tag{21}$$

Assuming that:

$$a_{i|l,k}^{0,1} = P\left(q_t^{0,1} = i | q_t^{0,1} = k\right)P\left(q_t^{0,1} = i | q_t^{1,0} = l\right), \tag{22}$$

the re-estimation of the transition probabilities may be simplified. For example:

$$P\left(q_t^0 = i | q_t^1 = k\right) \tag{23}$$

may be estimated as:

$$P\left(q_t^0 = i | q_t^1 = k\right) = \frac{\Sigma_r\frac{1}{P_r}\Sigma_t\Sigma_j\Sigma_l\alpha_t^r(k,l)\,a_{i,j|k,l}b_t^0(i)\,b_t^1(k)\,\beta_{t+1}^r(i,j)}{\Sigma_r\frac{1}{P_r}\Sigma_t\Sigma_j\Sigma_l\alpha_t^r(k,l)\,\beta_t^r(k,l)}. \tag{24}$$

The transitions from a non-emitting entry state i to any pair of audio-visual states (k,l) is given by:

$$a_{i|k,l} = \frac{1}{R}\sum_r\frac{1}{P_r}\alpha_1^r(k,l)\,\beta_1^r(k,l) \tag{25}$$

and the transitions from a state pair (k,l) to the exit non-emitting exit state o are given by:

$$a_{k,l|o} = \frac{\Sigma_r\frac{1}{P_r}\alpha_T^r(k,l)\,\beta_T^r(k,l)}{\Sigma_r\frac{1}{P_r}\Sigma_t\alpha_t^r(k,l)\,\beta_t^r(k,l)}$$

CHMM method is described based on the work of [14].

The maximum likelihood (ML) training of the dynamic Bayesian networks in general and of the coupled HMMs in

particular, is a well understood technique. However, the iterative maximum likelihood estimation of the parameters only converges to a local optimum, making the choice of the initial parameters of the model a critical issue. In [16] there is presented an efficient method for the initialization of the ML training that uses a Viterbi algorithm derived for the coupled HMM. The Viterbi algorithm determines the optimal sequence of states for the coupled nodes of the audio and video streams that maximizes the observation likelihood. The following steps describe the Viterbi algorithm for the two stream coupled HMM used in our audio-visual system. An extension to a multi-stream coupled HMM is straightforward [16]:

– Initialization:

$$\delta_0(i,j) = \pi_0^a(i)\,\pi_0^v(j)\,b_t^a(i)\,b_t^v(j), \tag{26}$$

$$\psi_0(i,j) = 0. \tag{27}$$

– Recursion:

$$\delta_t(i,j) = \max_{k,l}\left\{\delta_{t-1}(k,l)\,a_{i|k,l}a_{j|k,l}\right\} b_t^a(k)\,b_t^v(l), \tag{28}$$

$$\psi_t(i,j) = \arg\max_{k,l}\left\{\delta_{t-1}(k,l)\,a_{i|k,l}a_{j|k,l}\right\}. \tag{29}$$

– Termination

$$P = \max_{i,j}\left\{\delta_T(i,j)\right\}, \tag{30}$$

$$\left\{q_T^a, q_T^b\right\} = \arg\max_{i,j}\left\{\delta_T(i,j)\right\}. \tag{31}$$

– Backtracking

$$\left\{q_t^a, q_t^v\right\} = \psi_{t+1}\left(q_{t+1}^a, q_{t+1}^v\right). \tag{32}$$

The segmental K means the algorithm for the coupled HMMs is described in the following steps:

Step 1. For each training observation sequence r, the data in each stream is uniformly segmented according to the number of states of the coupled nodes and an initial state sequence for the coupled nodes:

$$Q = q_{r,0}^{a,v}, ..., q_{r,t}^{a,v}, ..., q_{r,T-1}^{a,v} \tag{33}$$

is obtained. For each state i of the coupled nodes in stream c the mixture segmentation of the data assigned to it is obtained using the K-means algorithm with M_i^c clusters. Consequently the sequence of mixture components:

$$S = s_{0,r}^{a,v}, ..., s_{r,t}^{a,v}, ..., s_{r,T-1}^{a,v} \tag{34}$$

for the mixtures nodes is obtained.

Step 2. The new parameters of the model are estimated from the segmented data.

$$\mu_{i,m}^{a,v} = \frac{\Sigma_{r,t}\gamma_{r,t}^{a,v}(i,m)\,O_t^{a,v}}{\Sigma_{r,t}\gamma_{r,t}^{a,v}(i,m)}, \tag{35}$$

$$\sigma_{i,m}^{2\;a,v} = \frac{\Sigma_{r,t}\gamma_{r,t}^{a,v}(i,m)\left(O_t^{a,v}-\mu_{i,m}^{a,v}\right)\left(O_t^{a,v}-\mu_{i,m}^{a,v}\right)^T}{\Sigma_{r,t}\gamma_{r,t}^{a,v}(i,m)}, \tag{36}$$

$$w_{i,m}^{a,v} = \frac{\Sigma_{r,t}\gamma_{r,t}^{a,v}(i,m)}{\Sigma_{r,t}\Sigma_m\gamma_{r,t}^{a,v}(i,m)}, \tag{37}$$

$$a_{i|k,l}^{a,v} = \frac{\Sigma_{r,t}\varepsilon_{r,t}^{a,v}(i,k,l)}{\Sigma_{r,t}\Sigma_k\Sigma_l\varepsilon_{r,t}^{a,v}(i,k,l)}, \tag{38}$$

where

$$\gamma_{r,t}^{a,v}(i,m) = \begin{cases} 1, & \text{if } \; q_{r,t}^{a,v} = i,\; s_{r,t}^{a,v} = m \\ 0, & \text{otherwise} \end{cases} \tag{39}$$

$$\varepsilon_{r,t}^{a,v}(i,k,l) = \begin{cases} 1, & \text{if } \; q_{r,t}^{a,v} = i \\ & \quad q_{r,t-1}^{a} = k, \;\; q_{r,t-1}^{v} = l \\ 0, & \text{otherwise} \end{cases} \tag{40}$$

Step 3. At consecutive iteration the optimal state sequence $Q$of the coupled nodes is obtained using the Viterbi algorithm (Eqs. (26)–(32)). The sequence of mixture component S is obtained by selecting at each moment t the mixture $s_{r,t}^{a,v}$ such that:

$$s_{r,t}^{a,v} = \max_{m=1,...,M_i^{a,v}} P\left(O_t^{a,v}|q_{r,t}^{a,v} = i,m\right). \tag{41}$$

Step 4. The iterations in steps 2–4 are repeated until the difference between the observation probabilities of the training sequences at consecutive iterations falls below the convergence threshold.

The word recognition is carried out via the computation of the Viterbi algorithm (Eqs. (26)–(32)) for the parameters of all the word models in the database. The parameters of the CHMM corresponding to each word in the database are obtained in the training stage using clear audio signals (SNR = 20 db). In the recognition stage the influence of the audio and visual streams is weighted based on the relative reliability of the audio and visual features, for different levels of the acoustic noise. Formally the observation probability at time t for the observation vector $O_t^{a,v}$ becomes:

$$\widetilde{b}_t^{a,v}(i) = b_t\left(O_t^{a,v}|q_t^{a,v} = i\right)^{\alpha_{a,v}}, \tag{42}$$

where $\alpha_a + \alpha_v = 1$ and $\alpha_a,\, \alpha_v \geq 0$ are the exponents of the audio and video streams. The values of α_a, α_v corresponding to a specific acoustic SNR level are obtained experimentally to maximize the average recognition rate.

Learning the system based on the Viterbi algorithm is described based on the work of [16].

5. The method of extraction of video features

The paper proposes a method of speech recognition, based on the characteristics of the audio and visual signal. As the video information of speech, we accepted an observation vector, created in the process of feature extraction video.

The first step in process of creating video observation speech vectors is the location of the user's face in a video. In this work, there is used a detection method, based on Haar-like features – for face localization. In a Haar-like feature approach, feature values are obtained by summing up the values of pixels in each region of a face image and weighting and then summing up the regional sums, instead of directly using the values of the pixels of the face image. Haar-like feature is a linear combination of the intensity sum of pixels (several rectangular regions), which use two different rectangular regions. After determining the coordinates of vertices

rectangular mask, there is created a new video sequence of statements containing the limited area of the image to the user's face – from the original sequence of frames.

In order to point an area of image containing the user's mouth can be used to determine the coordinates of eyes. For this reason Gradient Method and Integral Projection (GMIP) [17] is applied to find horizontal and vertical lines of eyes. Dependencies used to determine the boundaries of the mouth and the results of determination of the mouth area based on the position of eyes, are shown in Fig. 2.

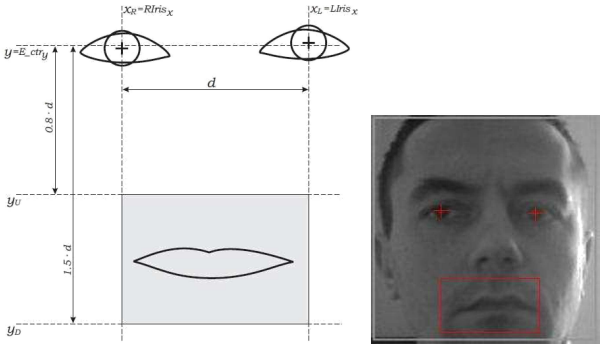

Fig. 2. Dependencies used to determine the boundaries of the mouth and the results of determination of the mouth area based on the position of eyes

The next step in the creation of video observation vectors of speech is to locate the edge of user's mouth. The paper proposes a method based on a specific colour and shape of lips [18]. In this method, the localization process of lip corners is realized on a colour image. Because lip colour is so specific, it is possible to manipulate the various components of the RGB in order to determine isolated border between the lips and the rest of the face, by thresholding. In this way, one sets the values of the pixels corresponding to the specific colour of the lips. The method of operation on RGB describes the following relationship:

$$lips\_region = \begin{cases} \dfrac{B}{G} - 1 < T1 \\ \dfrac{R}{G} - \dfrac{B}{G} < T2 \\ \dfrac{R}{G} - 1 < T3 \end{cases}, \tag{43}$$

where $T1$, $T2$ and $T3$ are empirically chosen thresholding.

The audio speech may distinguish all phonemes. During speaking in the video it happens that – for various phonemes – lips have a very similar look and layout. Some phonemes can be distinguished in a video speech by observation whether the tongue can be seen explicitly between the teeth. For example, the mouth when speaking in Polish phonemes "a" and "l" are arranged very similarly, but for the phoneme "l" tongue may be seen near the teeth. Similarly, no apparent tongue pronounces phonemes "i" and "j", and with the tongue for phonemes "t" and "d". On this basis, the system introduces the appropriate weights for the phonemes of the observed tongue in order to improve the differentiation of individual phonemes of video. The system looks for an area of colour similar to the tongue or mouth, which appears near the upper teeth. After finding such an area, the system checks the brightness level grayscale, and if the level is larger than a specified threshold, it means the area as a visible tongue.

The system is based on CHMM. For the CHMM model, the input signal has to be introduced as a vector of observations. For each frame, based on the coordinates of characteristic points, there is assigned a symbol that best describes the characteristics of that frame. The proposed method for encoding frames uses a simplified method that uses the location of each characteristic point of the straight line defined by the corners of the mouth. One calculates the sum of relative distances $m$from all points of a straight line, defined by the corners of the mouth, for each frame. There are adopted 16 characteristic points, so each of the calculated relationships is divided by 16. The value of obtained sum, multiplied by 100, is in the range from 11 to 60, obviously when properly located in the characteristic points on the outer edges of the lips:

$$y = \frac{\sum_{i=1}^{N} \frac{m_i}{d}}{N} \cdot 100, \tag{44}$$

where: N – is the number of points.

Video speech encoding method shown in Fig. 3.

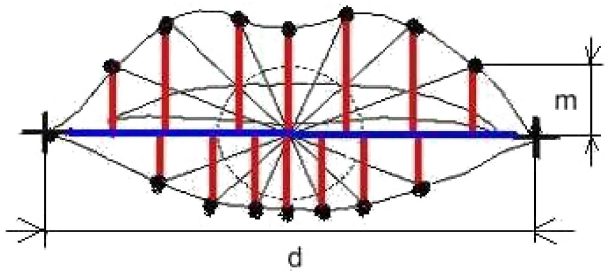

Fig. 3. Video speech encoding method

It is assumed that the resulting symbols should be in the range from 1 to 50, so the minimum value of the code for each user must be specified and on this basis code values to the objective range must be reduced. After an analysis of the video speech encoding, it was noted that the code for most users at the mouth closed revolved around the number 11. Number of distinct phonemes with a visible tongue is small, so one may use the range from 1 to 10 for encoding such phonemes and the remaining range from 11 to 50 may be used to encode the phonemes without a visible tongue.

6. Assumptions of the system of audio-visual speech recognition

In the work, our method of audio-visual speech recognition, called *AV_Mowa_PL*, was proposed to limit the negative influence of external factors on audio speech. The method based on hidden Markov models, was worked up for recognition of Polish audio-visual speech. Novel peculiarity of the method was the use of a data vector, where the audio and visual signals

of Polish speech were joined. The method may be proposed also to identify the speaker.

The audio-visual speech recognition is based on the extraction of recording features of audio and features of video. There are analysed, in such a system, separately video and audio channels. That makes a proper fusion of designated features by using CHMM.

Both the signal sources were analysed separately for the creation of the observation vector, containing the necessary characteristics of the source audio and video. Peculiarity of this method depends on used observations vectors of the audio and video in speech as teaching data for the same CHMM. Because in the system presented in this paper, the audio visual CHMM allows for asynchrony in the audio and visual states (skip the step of synchronization) but forces them to be synchronized at the model boundaries (adopted such a selection of frequency recording audio and video signals) allows to obtain the same length of observation vectors. Such requirements are met by application of the audio signal sampling frequency 8000 samples per second with a delay of 80 frames of samples, and the use of frequencies for the picture 50 frames per second. Diagram of the operation method using CHMM is shown in Fig. 4.

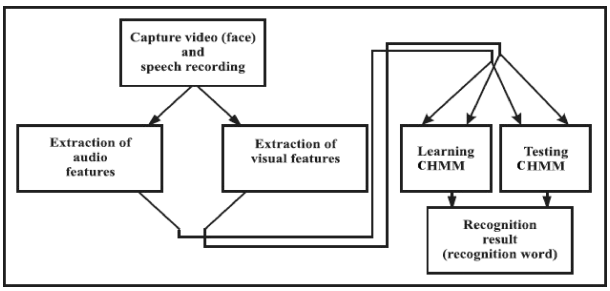

Fig. 4. Diagram of the operation method using CHMM

The learning process uses an approach based on stochastic models, CHMM used for modeling time series. Learning CHMM model is the best fit value of its parameters. Preceding the learning process, it is necessary to determine the topology of the model. The proper selection of the number of states influences the accuracy and the speed of learning. The learning process is to estimate the parameters of the model, for a given learning sequence of observations. The given observation sequence consists of several repetitions of the same speech and audio encoded in the form of observation symbols. In isolated word the recognition system was used for each of your separate model described in the appropriate grammatical transcription. The number of models corresponds to the number of words contained in the dictionary of system. Adding new words to the dictionary associated with the creation of a new model, while in the learning process only the parameters of this new model are determined, the parameters of the other models do not change. Additionally there was assumed that all models were the same size. For a new user of the system, it is necessary to create a new set of models for each expression. One may learn from the new models already in place, but then lose their value for the previous user.

In the process of recognition it is made a similar analysis and coding, as in the case of learning the system. Recognized word represented by a observation vector is compared with all the models in the CHMM system. Recognition consists of determining the likelihood of generating the input sequence of observations by the model because each model CHMM may be seen as a generator sequence of observations. Recognition determines also the maximum likelihood model that was trained on the data most similar to the recognized word. The result of recognition is the equivalent of a winning transcription grammatical model.

The method of audio-visual recognition of the Polish language is implemented in the system to control the industrial camera, with voice and lip movements – to test it and select the optimum parameters. Due to the limited number of commands to control the camera, the system allows to teach any number of words, regardless of the reference control. This approach allows to test a large number of commands, while retaining the ability to control camera movement using the selected voice commands. View of such a system is shown in Fig. 5.

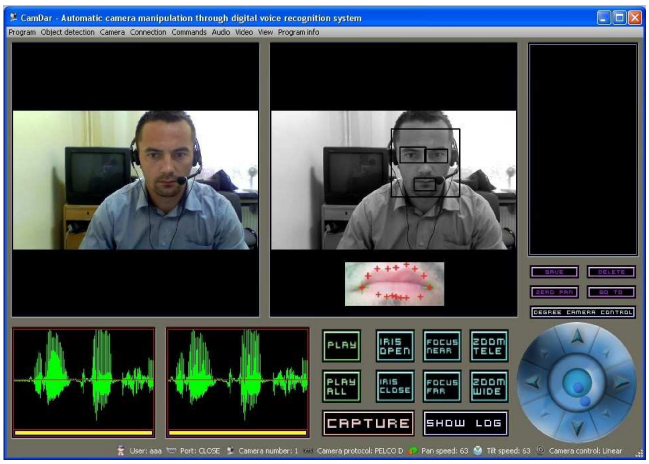

Fig. 5. View the system to control the camera movement through audio-visual speech

7. Experimental results

Some experimental research began with the selection of the optimal code-book size and number of hidden states of CHMM. It was assumed that an appropriate choice of these parameters would allow for the proper conduct of the main experiment. The study realized on undisturbed conditions for audio speech, containing about fifty different commands.

When choosing the size of the code books, it was suggested by the fact that in Polish there are 37 distinct phonemes. Besides the phonemes, all the possible transitions of phonemes are taken into consideration, but then the code-book size increases to over a thousand values. Therefore, the study was done for the code-book size close to the number of phonemes of the Polish language, and more specifically for the size specified respectively: 32, 37, 64.128, 256

The experience of choosing the number of codes in the code book was carried out for different numbers of states of

CHMM models. Adopted by the number of states equal to 5, 8, 10 and 15. During the tests, using words with different numbers of phonemes, and it was necessary to choose the number of states that was appropriate for the given command reference. Error rates were presented in the form of incorrect rejection (False Rejection Rate, FRR) and false acceptance (False Acceptance Rate, FAR) for the average of all. Erroneous rejection means that the statement was not recognized, while false acceptance means that the utterance was recognized correctly. The results of the first experiment are shown in Table 1.

Table 1
Results of the selection experiment the number of codes in the book code and numbers of states of CHMM models

Number of states of CHMM	Codebook size	FAR [%]	FRR [%]	Recognition performance [%]
5	32	4	8	88
	37	4	4	92
	64	6	6	88
	128	8	6	86
	256	8	8	84
8	32	4	0	96
	37	**0**	**0**	**100**
	64	0	2	98
	128	4	4	92
	256	6	8	86
10	32	4	10	86
	37	2	4	94
	64	4	10	86
	128	6	8	86
	256	6	10	84
15	32	4	10	86
	37	4	6	90
	64	4	12	84
	128	4	14	82
	256	6	12	82

Then the effectiveness of only audio and audio-visual speech recognition was examined in conditions of distorted audio signal of speech. To carry out the research there was applied a set of two hundred commands, recorded at a frequency of 8000 samples per second (for audio) and 50 frames per second (for video). 40 different users were tested. In order to show the effectiveness of the method of Polish audio-video speech, experiments were performed for different levels of audio noise (at SNR of 20, 15, 10, 5, and 0 dB) and different number of states of CHMM. The same test was performed for the audio only speech recognition. Adopted at 20 dB SNR of at undisturbed conditions for sound recording. Checked, how it affects the efficiency of detection the number of states in a noisy. It was assumed equal number of states for audio and video speech for audio-visual speech recognition. Uses the number of states as in the earlier study, respectively: 5, 8, 10 and 15. Results of the second experiment are shown in Table 2.

Table 2
Recognition results of audio-video speech for different levels of noise and different numbers of states

Audio-visual speech recognition with use method based on CHMM					
Number of states of CHMM	Recognition Accuracy [%]				
	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB	SNR 0 dB
5	93.83	85.12	79.53	69.08	64.11
8	95.91	88.98	82.02	75.15	70.09
10	**96.12**	**89.76**	**83.33**	**77.11**	**71.32**
15	94.05	88.02	81.54	74.71	69.58
Audio speech recognition with use method based on HMM					
Number of states of HMM	Recognition Accuracy [%]				
	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB	SNR 0 dB
5	92.02	74.81	48.19	30.60	21.96
8	**96.02**	**77.35**	**52.30**	**37.84**	**29.73**
10	95.93	76.86	50.49	33.28	27.34
15	93.16	76.05	50.09	32.87	28.13

Many scientists in the world deal with the analysis of audio-visual speech. In their studies, they examine the various factors of processing audio-visual speech. Therefore, to compare the obtained results with those of other researchers, we chose only those works that were analysed in a similar way audio-visual recognition of speech. In order to compare the developed method with the popular methods of audio-visual recognition of the speech, developed by leading researchers in this field, we adopted similar conditions for noisy audio signal. Effectiveness compared with those of: [2] – *AV Combined*, [5] - *Audiovisual*, [7] – *AV-LSNR*, [9] – *AV-CHMM*, [15] – *AV-Concat, AV-HiLDA, AV-Enhanced, AV-MS-Joint*, in which the authors have adopted similar solutions to encode both signals, and the use of CHMM for learning and testing. Assumptions may differ in terms of quantity of the analysed words, different amounts of CHMM states and various means of fusion of audio and video signals. But the sense of studies was similar, so it was concluded that the comparison would be reliable. The results of comparing the level of recognition errors of audio-visual speech was showed in Table 3.

Table 3
The results of comparing the level of recognition errors of audio-visual speech for different methods

Method	Recognition Accuracy [%]				
	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB	SNR 0 dB
AV-Concat	88.37	80.66	73.95	66.36	53.73
AV-HiLDA	88.44	81.92	75.91	66.77	56.49
AV-Enhanced	87.28	79.84	70.28	56.88	41.04
AV-MS-Joint	88.63	82.52	77.08	69.97	59.11
AV-LSNR	93.13	88.26	83.05	78.64	71.27
AV-CHMM	98.56	90.08	85.09	75.23	70.51
Audiovisual	94.72	88.16	84.72	74.12	68.96
AV Combined	97.20	93.40	79.50	58.40	50.80
AV_Mowa_PL	**96.12**	**89.76**	**83.33**	**77.11**	**71.32**

8. Conclusions and future work

After the first part of the study there was found that the number of all possible combinations of transitions between phonemes were far too large in size of the dictionary. A study of the effectiveness of the recognition vocabulary size equal to respectively 256, 128, 64, 37 and 32. There appeared, for sizes larger than 37, the phenomenon of the same phoneme code using several completely different codes, which introduced the possibility of erroneous recognition of words presented in the form of the observation vector, where each observation corresponded to a single code from the code book. At size 32, there was a situation in which the various phonemes were encoded using the same code. Observing the recognition results for different sizes of code book, it was found that the best results could encode the signals using the 37 codes. By adopting such a code-book size dropped from the phenomenon of transition between phonemes.

In the second experiment, there was observed that an excessive number of states made it difficult to identify the correct words, while a too small number of states results in a misdiagnosis. The best results, based on the audio speech, obtained by the number of states equal eight HMM models. Analysing the obtained results, it may be assumed that the number of states should correspond to the number of phonemes of the word. But there was no need or possibility of using models with different numbers of states for different words, so be sure to choose the number of states of HMM models respectively for the base words in the system. The best for a specific database commands take the number of states equal to the average number of phonemes per one command from the database. For the analysed database commands for audio-visual speech, the best results were obtained for the ten states of CHMM, while for the same audio speech, the best results were obtained when eight states of HMM. Results of the second experiment are shown in Fig. 6.

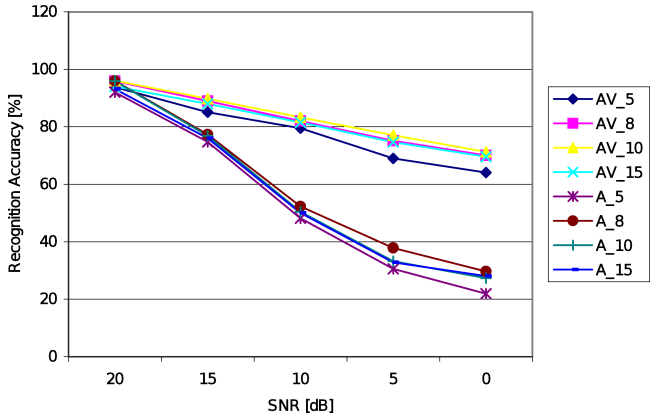

Fig. 6. Recognition results of audio and audio-visual speech,with varying degrees of noisy audio signal and different numbers of states (numbers with names indicate the number of used states of HMM and CHMM)

The third experiment showed that the method obtained similar or better results to other existing audio-visual speech recognition methods, published in scientific literature. Figure 7 shows results of comparing the level of recognition errors of audio-visual speech for different methods.

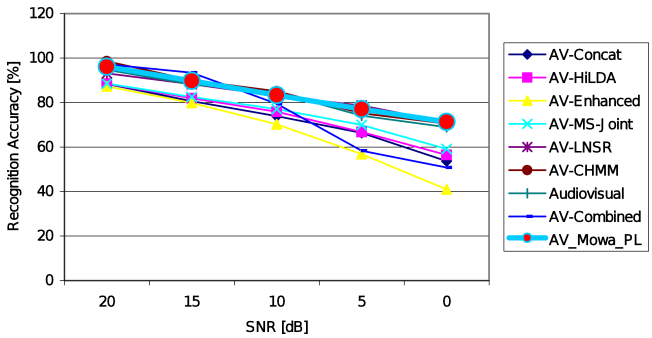

Fig. 7. The results of comparing the level of recognition errors of audio-visual speech for different methods

Based on the tests there was shown that the method of Polish audio-visual speech recognition worked properly and it could work in the systems in the real world. Test results show the accuracy of speech recognition that a large impact on the proper identification was affected, disturbed or not, the environment. The results showed also that this method should be developed. There are plans to expand the method of automatic detection of the position of the tongue, for each of the spoken video phonemes. Further work will also build a system for Polish speech recognition, based on an analysis of individual phonemes. Such an approach would allow for continuous speech recognition. The method of audio-visual recognition of Polish speech was used in the system to control the camera movement using voice commands. To increase the efficiency of the method FPGA (Field Programmable Gate Array) can be used. As a consequence the system is enabled to work properly in a real time as well as the hardware level is supported [19].

An advantage of the proposed method is the satisfactory effectiveness created by the lip-tracking procedures, and the simplicity and functionality by the proposed methods, which fuse together the audio and visual signals. A decisively lower level of mistakes was obtained in audio-visual speech recognition, and speaker identification, in comparison to only audio speech, particularly in facilities, where the audio signal is strongly disrupted.

REFERENCES

[1] T. Saitoh, K. Morishita, and Konishi, "Analysis of efficient lip reading method for various languages", *Pattern Recognition 19th Int. Conf. on ICPR* 1, 1–4 (2008).

[2] T. Yoshinaga, S. Tamura, K. Iwano, and S. Furui, "Audio-visual speech recognition using new lip features extracted from side-face images", *Proc. Auditory Visual Speech Processing (AVSP)* 1, 117–120 (2003).

[3] Jongju Shin, Jin Lee, and Daijin Kim, "Real-time lip reading system for isolated Korean word recognition", *Pattern Recognition* 1, 559–571 (2011).

[4] W.J. Ma, X. Zhou, L.A. Ross, J.J. Foxe, and L.C. Parra, "Lip reading aids word recognition most in moderate noise: a Bayesian explanation using high- dimensional feature space", *PLoS ONE* 4 (3), 1–14 (2009).

[5] M. Hasegawa-Johnson, K. Livescu, P. Lal, and K. Saenko, "Audiovisual speech recognition with articulator positions as hidden variables", *Proc. Int. Congress of Phonetic Sciences (ICPhS)* 1, CD-ROM (2007).

[6] Y. Kim, S. Kang, and S. Jung, "Design and implementation of a lip reading system in smart phone environment", *Proc. 10th IEEE Int. Conf. on Information Reuse and Integration* 1, 101–104 (2009).

[7] X. Shao and J. Barker, "Stream weight estimation for multistream audio–visual speech recognition in a multispeaker environment", *Speech Communication* 50, 337–353 (2008).

[8] L. Jong-Seok and P. Cheol Hoon, "Robust audio-visual speech recognition based on late integration", *IEEE Trans. on Multimedia* 10 (5), 767–779 (2008).

[9] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition", *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* 1, CD-ROM (2002).

[10] L. Rabiner and B.H.Yuang, "Fundamentals of speech recognition", *Prentice Hall Signal Processing Series* 1, CD-ROM (1993).

[11] A.M. Wisniewski, "Implicit Markov models in speech recognition", *Bulletin IAiR WAT* 7, CD-ROM (1997), (in Polish).

[12] G. Demenko, B. Mobius, and K. Klessa, "Implementation of Polish speech synthesis for the BOSS system", *Bull Pol Ac.: Tech.* 58 (3), 371–376 (2010).

[13] M.N.N. Kaynak, Q. Zhi, A.D. Cheok, K. Sengupta, and K.C. Chung, "Audio - visual modeling for bimodal speech recognition", *Proc. Int. Fuzzy Systems Conf.* 1, CD-ROM (2001).

[14] X. Liu, Y. Zhao, X. Pi, L. Liang, and A.V. Nefian, "Audio-visual continuous speech recognition using a coupled hidden Markov model", *ICSLP-2002* 1, 213–216 (2002).

[15] C. Neti, G. Potamianos, J. Luttin, I. Mattews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio visual speech-recognition", *2000 Final Report* 1, CD-ROM (2000).

[16] A. V. Nefian, L. Liang, X. Pi, X. Liu, and C. Mao, "A coupled hidden Markov model for audio-visual speech recognition", *Int. Conf. on Acoustics, Speech and Signal Processing* 1, CD-ROM (2002).

[17] G. Kuchariev and A. Kuzmiski, *Biometric Technology. Part 1: Methods for Face Recognition*, Technical University of Szczecin, Szczecin, 2003, (in Polish).

[18] M. Choras, "Human lips as emerging biometrics modality", *Image Analysis and Recognition, ICIAR 2008* 1, 994–1003 (2008).

[19] S. Szczepanski, M. Wojcikowski, B. Pankiewycz, M. Klosowski, and R. Zaglewski, "FPGA and ASIC implementation of the algorithm for traffic monitoring in urban areas", *Bull. Pol. Ac.: Tech.* 59 (2), 137–140 (2011).