УДК 004.8

Ю.В. КРАК, А.В. БАРМАК, А.С. ТЕРНОВ

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ДЛЯ АВТОМАТИЧЕСКОГО ЧТЕНИЯ ПО ГУБАМ УКРАИНСКОЙ РЕЧИ

Предложена информационная технология для автоматического чтения по губам украинской речи. Для проведения эксперимента было создано программное обеспечение для локализации лица на фотографии, выделения множества точек для контура губ, разложения характеристического вектора по базису. Проведенные исследования подтвердили работоспособность предложенной технологии.

Введение. Одна из проблем «восприятия и понимания устной речи» [1] людьми с проблемами слуха при общении с другими людьми – умение распознавать разговорный язык по губам. С этой точки зрения задача распознавания по губам является альтернативой языкового общения для людей с недостатками слуха. Развитие направления автоматического чтения по губам также поможет усовершенствовать показатели существующих систем распознавания речи благодаря получению дополнительного независимого канала информации и создать учебные программы для слышащих людей с целью улучшения артикуляции губ при проговаривании слов для более точного распознавания.

В контексте изучения данной проблемы можно выделить фундаментальные исследования под руководством И.К. Билодида [2] по тематике артикуляторных особенностей при проговаривании в современной украинской литературе и языке, а также результаты работ В.И Бельтюкова по схожей проблематике для русского языка [3], состоящие в определении возможности визуальной идентификации фонем русского языка. Из этих и других исследований [4] можно сделать вывод о том, что визуальный алфавит существенно не полный. В нем нет однозначного соответствия между произнесенной фонемой и ее визуальным отображением, что приводит к снижению возможности зрительного восприятия речи. Однако описанные сложности не исключают возможности создания

системы обучения правильной артикуляции для облегчения зрительного восприятия устной речи людьми с нарушением слуха.

На сегодняшний день существует много публикаций, обзорных статей и диссертационных работ по данной тематике [5, 6]. Однако остается много нерешенных вопросов, связанных как с качеством распознавания, так и с фонетическими и грамматическими особенностями конкретных языков. В частности, для украинского языка требуется адаптация существующих методов и подходов.

Основными ключевыми моментами в решении поставленных задач аудиовизуального распознавания речи в контексте получения параметров визуальной компоненты динамики речевого процесса являются вариации методов по определению положения губ на изображении, способы получения динамических или геометрических характеристик изменения мимики губ, а также использование различных подходов к анализу полученных характеристик.

Обзор существующих результатов показал перспективность развития исследований в данном направлении. Для украинского языка необходимо адаптировать существующие методы и подходы к задаче чтения по губам; разработать новые системы обучения правильной артикуляции; изучить фонетические и морфологические особенности украинской речи и визуальную составляющую артикуляторных процессов динамики произношения.

Проведенный анализ существующих результатов и работ по данной тематике определил направление исследований и постановку задачи.

Постановка задачи. Необходимо синтезировать математическую модель, в рамках которой возможно выявить различия положений губ конкретного человека для построения системы обучения правильной артикуляции при воспроизведении слов украинской речи. Математическая модель должна включать в себя реализацию следующих возможностей:

- создание визуального алфавита украинского языка для конкретного человека;
- анализ визуального отображения фонем украинской речи для конкретного человека (в рамках созданного для него визуального алфавита);
- возможность применения полученных результатов для любой группы людей.

Математическая модель. Для синтеза математической модели предлагается перейти от пространства фотографических изображений лица человека в процессе проговаривания к векторному пространству характеристических параметров. Такой переход предлагается осуществить в несколько этапов.

1. Выделение на изображении внутреннего контура губ:

$$\mathrm{Im}L \to D, \qquad (1)$$

где $\mathrm{Im}L = \{I_k : I_k \in FSV\}$ – упорядоченное множество ключевых кадров видеопотока FSV (Face Speech Video), полученного при съемке процесса проговаривания слов украинского языка ($k = \overline{1,N}$ – порядковый индекс кадра в выбранной последовательности, где N – количество ключевых кадров);

– $I_k = \left\{ col_{ij}^k \right\}_{i,j=1}^{m,n}$, $i = 1,...,m;\ j = 1,...,n$ – изображение размером $m \times n$ лица с мимическим положением губ при проговаривании слов украинского языка, где m и n – соответственно длина и ширина изображения I_k;

– $col_{ij}^k = I_k(i, j)$ – цвет пикселя в системе RGB с координатами (i, j) на изображении I_k;

– $D = \left\{ D_k : D_k = \{ d_{top}^k, d_{bot}^k \} \right\}$ – множество контуров губ, где D_k – пара точечных кривых – контуров губ (верхний d_{top}^k и нижний d_{bot}^k) для k-го кадра.

2. Аппроксимация полученной точечной кривой внутреннего контура губ с помощью неравномерных базисных сплайнов (NURBS) – получение вектора характеристических признаков:

$$D \to P, \tag{2}$$

где $P = \{ v_k : v_k^i \in H, i = \overline{1, M} \}$ – пространство характеристических признаков; H – характеристические признаки объекта исследования; v_k – характеристический вектор, v_k^i – его координаты; M – размерность пространства P.

Выделение на изображении контура губ. Для выделения на изображении контура губ используется подход [7, 8]:

1-й шаг. Выделение на изображении I_1 области губ D_{mouth}^1 по алгоритму, в котором происходит нормирование изображения по расстоянию между центрами зрачков и ориентация по линии глаз. На остальных изображениях при построении D_{mouth}^1, $i = \overline{2, N}$ осуществляется только коррекция найденной на первом шаге области.

2-й шаг. Построение внутренних контуров d_{top} d_{bot} методами оконтуривания и цветовой сегментации в областях D_{mouth}^1, $i = \overline{2, N}$ (рис. 1).

Аппроксимация внутреннего контура губ. Переход $D \to P$ осуществляется с помощью следующих шагов:

1-й шаг. Выполняется предварительное сглаживание и выравнивание для получения контуров $\overline{d_{top}}$ и $\overline{d_{bot}}$.

2-й шаг. При нормировке точки $\overline{d_{top}^i}(x, y)$, где $x, y \in R$, контуры $\overline{d_{top}}$ переводятся в точки $\overline{d_{top,[0,1]}^i}(x, y)$, где $x, y \in [0,1]$ контура $\overline{d_{top}^{[0,1]}}$.

3-й шаг. По множеству точек контура $\overline{d_{top}^{[0,1]}}$ строится аппроксимирующая NURBS-кривая [9]. $p_{top}(u)$, где

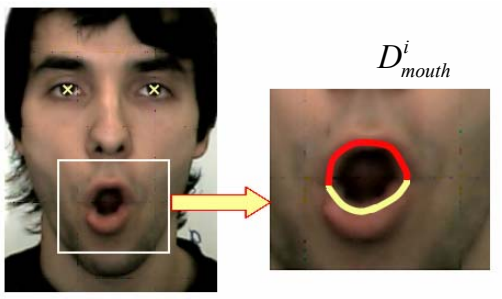


РИС. 1. Построение внутренних контуров губ d_{top} и d_{bot}

$$p(u)=\frac{\sum_{i=0}^{n} N_{i,d}(u)\,w_i\,p_i}{\sum_{i=0}^{n} N_{i,d}(u)\,w_i}, \tag{3}$$

где $p(u)=\left[x(u),\ y(u)\right]^T$ – функция, определенная на интервале $u_{\min}\le u\le u_{\max}$, такая, что является положительно гладкой и проходит, в некотором смысле, близко к опорным точкам $p_0,...,p_m$.

Функция $N_{id}(u)$ определяется с помощью рекурсивных функций Кокса-де Бура [10]:

$$\begin{aligned} N_{k,0} &= \begin{cases} 1, \text{ если } u_k \le u \le u_{k+1} \\ 0 \ - \quad \text{в} \quad \text{противном случае,} \end{cases} \\ N_{k,d} &= \frac{u-u_k}{u_{k+d}-u_k} N_{k,d-1}(u) + \frac{u_{k+d}-u}{u_{k+d+1}-u_{k+1}} N_{k+1,d-1}(u), \end{aligned} \tag{4}$$

где u_0, $u_1,...,u_n$ – последовательность узлов, такая что

$$u_{\min}=u_0\le u_1\le\cdots\le u_n=u_{\max}. \tag{5}$$

Воспользуемся простым свойством NURBS-кривых, которое следует из идентичности опорных точек (p_i) в однородной форме и равенства единице знаменателя. При $w_i=1$ NURBS-кривая (3) сводится к B-сплайн кривой. Учитывая, что при моделировании гибких шаблонов $w_i=1$, для упрощения аппроксимации можно перейти к B-сплайн кривым.

Задача B-сплайн аппроксимации – задача подгонки B-сплайн кривой с K опорными точками $p=\left[p_0,...,p_{K-1}\right]^T$ к точечной кривой $d=\left[d_0,...,d_{M-1}\right]^T$, где $M > K$ (обычно $M >> K$) для значений параметра $u_0,...,u_{M-1}$. Такая задача аппроксимации приводит к переопределенной системе линейных уравнений $N\cdot p=d$:

$$\begin{bmatrix} N_0(u_0) & \cdots & N_{K-1}(u_0) \\ N_0(u_1) & \cdots & N_{K-1}(u_1) \\ \vdots & \ddots & \vdots \\ N_0(u_{M-1}) & \cdots & N_{K-1}(u_{M-1}) \end{bmatrix} \cdot \begin{bmatrix} p_0 \\ \vdots \\ p_{K-1} \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{M-1} \end{bmatrix}, \quad (6)$$

где $N_i(u)$ – B-сплайн базисная функция (4).

Одним из способов получения решения переопределенной системы линейных уравнений (6) является следующий:

$$N^T N \cdot p = N^T \cdot d \text{, откуда } p = \left(N^T N\right)^{-1} \cdot N^T d \text{, при } \det\left(N^T N\right) > 0. \quad (7)$$

Для применения B-сплайн аппроксимации необходимо уметь получать на изображении точечные кривые $d = \left[d_0, \cdots, d_{M-1}\right]^T$ с тем, чтобы в дальнейшем использовать преобразования (7).

Таким образом, математической моделью мимических проявлений губ при проговаривании будет векторное пространство опорных точек NURBS-кривых:

$$P = \{v : v = (x_0^{p_{top}}, \cdots, x_{n_{top}-1}^{p_{top}}, x_0^{p_{bot}}, \cdots, x_{n_{bot}-1}^{p_{bot}}, y_0^{p_{top}}, \cdots, y_{n_{top}-1}^{p_{top}}, y_0^{p_{bot}}, \cdots, y_{n_{bot}-1}^{p_{bot}})\},$$
$$p_j^{p_{[top|bot]}} = (x_j^{p_{[top|bot]}}, y_j^{p_{[top|bot]}}), \quad j = \overline{0, n_{[top|bot]}-1}, \quad (8)$$

где $v \in P$ – вектор координат опорных точек $p_j^{p_{bot}}$ и $p_j^{p_{top}}$, аппроксимирующих NURBS кривых $p_{bot}(u)$, $p_{top}(u)$, а n_{bot} и n_{top} – количество контрольных точек для NURBS-кривых $p_{bot}(u)$, $p_{top}(u)$ соответственно.

В рамках предложенной математической модели, для получения векторного пространства будут проведены исследования различных базисов для определения наиболее подходящего, т. е. такого, разложение по которому даст лучший результат. Для этого предложена следующая технология.

Описание технологии. Схема технологии распознавания мимики при проговаривании слов на украинском языке показана на рис. 2.

Блок 1 отвечает за предварительную обработку входной визуальной информации и ее преобразование в пространство характеристических признаков (8).

Блок 2 содержит алгоритмы построения базиса пространства характеристических признаков и оценки его качества. На выходе строится базисная матрица векторного пространства характеристических признаков.

Блок 3 – происходит разложение вектора характеристических признаков, полученного для входного изображения по построенному базису.

Результатом работы технологии является вектор коэффициентов разложения, на основе которого принимается решение о соответствии входного изображения определенным базисным мимикам (при проговаривании слов украинского языка).

Построение базиса пространства характеристических признаков. Для построения базиса пространства на основе результатов исследований, приведенных в [2, 3], была выбрана группа L из 8 букв (звуков) украинского языка. В группу L входят буквы, которые имеют особенный артикуляционный портрет. Другими словами, те буквы, для которых мимические проявления (при их проговаривании) хорошо распознаются человеком.

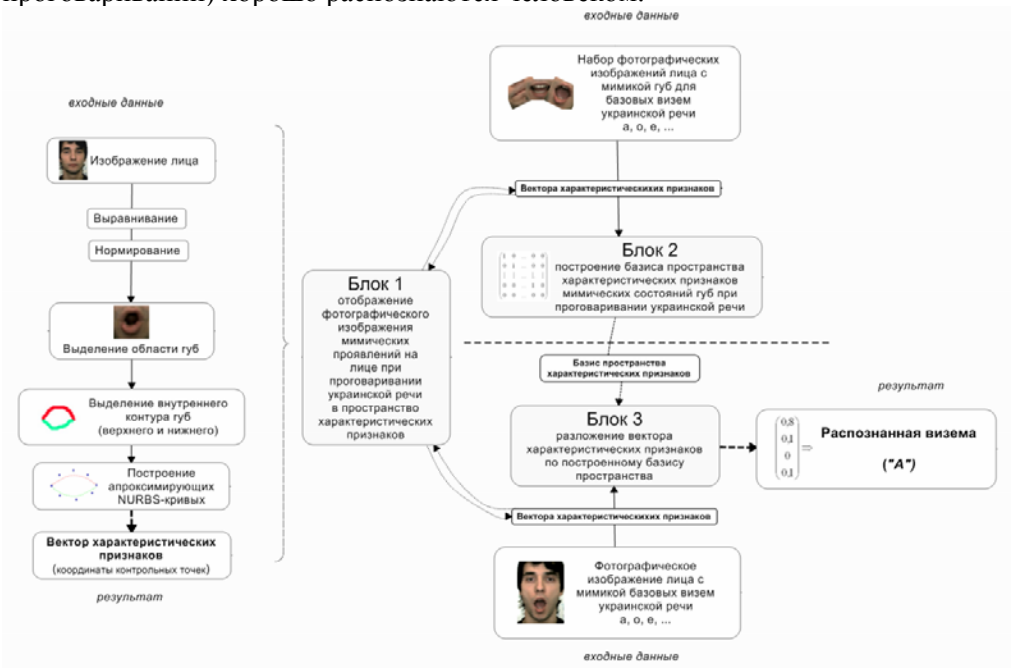


РИС. 2. Схема технологии распознавания мимики при проговаривании слов

Для компьютерного визуального представления мимических проявлений губ при проговаривании определенной буквы можно использовать следующие варианты:

1) один кадр (визема буквы [11, с. 331–349]) – кадр, на котором изображено максимальное отклонение губ от состояния спокойствия при проговаривании данной буквы (кадр I_3 на рис. 3);
2) последовательность из 3 кадров: средний кадр из интервала времени между состоянием спокойствия и виземой буквы, визема буквы, средний кадр из интервала времени между виземой буквы и состоянием спокойствия (кадры I_2, I_3, I_4 на рис. 3);
3) видеофрагмент с определенной частотой кадров (так чтобы количество кадров было больше 5); интервал времени между соседними состояниями спокойствия губ при проговаривании данной буквы.

Под состоянием спокойствия подразумевается положение губ, когда ничего не проговаривается.

В данном исследовании для построения базиса пространства использовался первый вариант.

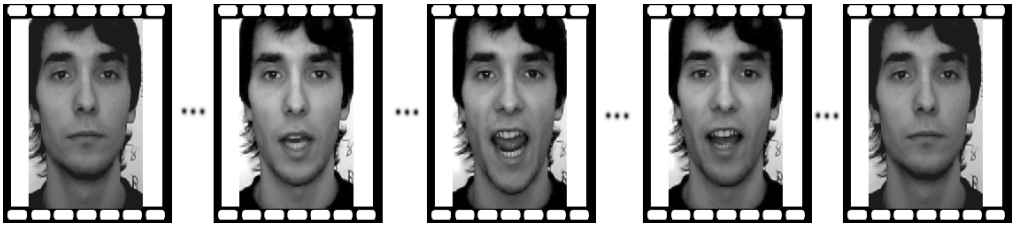


РИС. 3. Визуальное представление мимических проявлений губ при проговаривании определенной буквы Е

Для каждой буквы $l_i \in L$ формируется ее визема img_i, на основе которой строится вектор характеристических признаков. Набор полученных таким образом векторов будет составлять предположительный базис пространства (8).

Разложение произвольного вектора характеристических признаков по базису. Для разложения вектора характеристических признаков, построенного для входного изображения, по полученному базису рассматривается такая задача. Заданы матрица A размером $m \times n$ и b – вектор размерностью m. Необходимо найти все векторы x:

$$Ax = b. \tag{9}$$

Наиболее надежным методом для решения подобных задач является метод, основанный на матричной факторизации, – метод сингулярного разложения SVD [12]. На практике, для применения SVD, вводят порог τ близости к нулю сингулярных чисел, который представляет собой ошибки в исходных данных и ошибки вычислений. Тогда решение задачи (9) ищется таким образом:

где
$$x = A^{+}b = V\Sigma' U^{T}b, \tag{10}$$

$$\Sigma' = \begin{pmatrix} \sigma_1' & 0 & . & 0 \\ 0 & \sigma_1' & . & 0 \\ . & . & \ddots & . \\ 0 & 0 & . & \sigma_n' \end{pmatrix}, \ \sigma_j' = \begin{cases} \dfrac{1}{\sigma_j}, & \sigma_j \geq \tau \\ 0 - \text{другое.} \end{cases} \tag{11}$$

Решение наименьшей длины – приближенное решение задачи (9). Именно эта особенность – учитывать неточность начальных данных и приближенность вычислений считается преимуществом SVD подхода по сравнению с остальными.

Результаты моделирования. Для проведения эксперимента, на основе предложенной технологии, было создано оригинальное программное обеспечение (рис. 4) с такой функциональностью:

1) локализация области лица на фотографии;
2) нормирование и выравнивание изображения лица человека по расстоянию между центрами зрачков;
3) выделение множества точек для верхнего и нижнего контуров губ;
4) разложение характеристического вектора по базису.

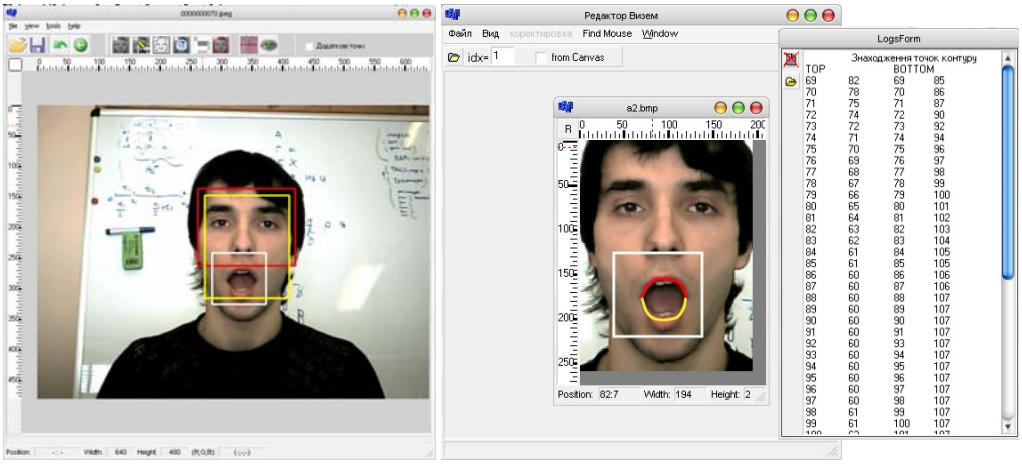
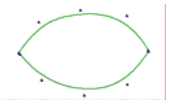
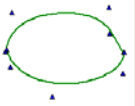


РИС. 4. Программа поиска точек контура губ

Исследования проводились на выборке из шести гласных визем ([i], [и], [e], [y], [o], [a]). Матрица A была получена на основе построения характеристических векторов базисной выборки. Под виземой понималось следующее: один кадр или визема буквы (см. рис. 3) .

В исследованиях при аппроксимации верхнего и нижнего контуров губ использовались NURBS-кривые с пятью опорными точками. В соответствии с таким выбором векторы пространства характеристических признаков имели размерность 20 (табл. 1).

ТАБЛИЦА 1

Виземы	Изображения лица	Контуры губ	NURBS-кривые	Вектор характеристических признаков
a				(0.171,0.274,0.486,0.725,0.835,0.178,0.287,0.507,0.726,0.836,0.363,0.603,0.694,0.650,0.384,0.614,0.417,0.296,0.382,0.639)
o				(0.369,0.388,0.678,0.680,0.720,0.391,0.400,0.526,0.734,0.740,0.403,0.634,0.623,0.527,0.431,0.626,0.464,0.353,0.451,0.550)

Для возможности использования метода наименьших квадратов без SVD разложения проверяется следующее неравенство:

$$val = \det(A^T A) >> 0 . \qquad (12)$$

В проведенных исследованиях $val = 0{,}002058$, поэтому полученная базисная матрица оказалась не хорошо обусловленной, вследствие чего был использован метод SVD для разложения по построенному базису пространства характеристических признаков.

На распознавание отправлялись гласные виземы, которые не использовались при построении базиса. Коэффициенты разложения вычислялись по (10). Матрица $\sum$ в экспериментах имела следующие диагональные элементы: $5{,}278949$; 0,624701; 0,487618; 0,219732; 0,128405.

Вектор характеристических признаков для изображения (виземы):

v_{a+} =(0.1908, 0.2961, 0.5066, 0.7171, 0.8224, 0.1908, 0.2961, 0.5066, 0.7171, 0.8224, 0.3993, 0.5291, 0.6848, 0.5486, 0.4019, 0.6068, 0.4486, 0.3687, 0.4311, 0.6021);

v_{e+} =(0.0921, 0.2303, 0.5066, 0.7829, 0.9211, 0.0921, 0.2303, 0.5066, 0.7829, 0.9211, 0.4316, 0.6254, 0.5674, 0.5932, 0.4519, 0.6559, 0.4039, 0.2651, 0.3667, 0.6763).

Результаты распознавания (разложения) для тестовых изображений приведены в табл. 2.

ТАБЛИЦА 2

	x(*a*+)	*x*(*e*+)	*x*(o+)
a	0,6	0,1	0,1
o	0,0	0,0	0,8
y	0,3	0,0	0,03
e	0,0	0,6	0,02
i,u	0,1	0,3	0,05

Отсюда видно, что наибольший коэффициент разложения соответствует базисной виземе.

Заключение. Проведенные исследования подтвердили работоспособность предложенной технологии.

Данная технология, кроме вывода о принадлежности исследуемой виземы классу соответствующей базисной, позволяет делать структурный анализ входных данных (изображений губ человека при произнесении слов украинского языка), сутью которого можно считать определение относительного вклада каждой базисной виземы.

Дальнейшие исследования будут направлены на

- построение более качественного базиса пространства характеристических признаков, благодаря расширению базисной выборки на все классы визем украинского языка;
- использование предложенного подхода для анализа артикуляции мимики при произнесении слов украинского языка.

Ю.В. Крак, *О.В. Бармак*, *А.С. Тернов*

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ДЛЯ АВТОМАТИЧНОГО ЧИТАННЯ ПО ГУБАХ УКРАЇНСЬКОЇ МОВИ

Запропоновано інформаційну технологію для автоматичного читання по губах української мови. Для проведення експерименту було створене програмне забезпечення для локалізації обличчя на фотографії, виділення множини точок контуру губ, розкладу характеристичного вектора за базисом. Проведені дослідження підтвердили працездатність технології.

Y. Krak, *O. Barmak*, *A. Ternov*

INFORMATION TECHNOLOGY DESIGNED FOR AUTOMATIC LIP READING FOR UKRAINIAN LANGUAGE

The information technology for automatic lip reading of Ukrainian language is proposed. The software for experimental use was created for face localization on the photo, for highlighting the set of lip contour points, and for characteristic vector expansion. The applicability of the technology was proved by numerous experiments.

1. *Жук В.В.* Навчання мови і мовлення дітей з порушеннями слуху за новими програмами // Дефектологія. – 2007. – № 2. – С. 3–6.
2. *Білодід І.К.* Сучасна українська літературна мова / І.К. Білодід.: – К. Ін-т мовознавства ім. О.О. Потебні; Наук. думка, 1969. – 435 с.
3. *Бельтюков В.И.* Чтение с губ фонетических элементов речи. – М.: Просвещение, 1967. – 143 с.
4. *Chan T.* HMM-based audio-visual speech recognition integrating geometric and appearance-based visual features // IEEE 2nd Workshop on Multimedia Signal Processing, Oct 3–5 2001. – P. 9–14.
5. *Audio-visual* speech recognition : Final Workshop 2000 Report / Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, Md, USA., 2000. – 84 p. http://www.idiap.ch/ftp/reports/2000/com00-02.pdf
6. *Sascha F.* Audiovisual speech: analysis, synthesis, perception and recognition // 16th International Congress of Phonetic Sci., Saarbrucken, August 2007. – P. 275–278.
7. *Soldatov S.* Lip Reading: Preparing Feature Vectors // Graphicon . – M., 2003. – P. 254–256.
8. *Крак Ю.В., Кривонос Ю.Г., Тернов А.С.* Локалізація і врахування особливостей обличчя людини для задачі розпізнавання за портретною фотографією // Штучний інтелект. – 2007. – № 3. – С. 229–236.
9. *Крак Ю.В., Бармак О.В., Єфімов Г.М.* Використання контурних моделей для побудови базису простору мімічних виразів емоцій // Штучний інтелект. – 2007. – № 4. – С. 288–296.
10. *Piegl Les*. Tiller Wayne The NURBS Book, and Edition. – Berlin, Germany: Springer-Verlag, 1996. – 645 p.
11. *Stork D.G., Hennecke M.E.* Speechreading by Humans and Machines. – Berlin: Springer, 1996. – 666 p.
12. *Форсайт Дж.* Машинные методы математических вычислений: Пер. с англ. Х.Д. Икрамова. – М.: Мир, 1980. – 277 с.

Получено 23.12.2008

Об авторах :

Крак Юрий Васильевич,
доктор физико-математических наук, профессор кафедры моделирования сложных систем Киевского национального университета имени Тараса Шевченко,
e-mail krak@unicyb.kiev.ua

Бармак Александр Владимирович,
кандидат технических наук, доцент, старший научный сотрудник Института кибернетики имени В.М. Глушкова НАН Украины,

Тернов Антон Сергеевич,
аспирант Института кибернетики имени В.М. Глушкова НАН Украины.