Чтение по губам в жестовой речи: синтез и анализ

Крак Ю.В.,

Тернов А.С.

В статье рассматривается технология анализа визуальной составляющей речевого процесса в видеопотоке. Данная технология позволяет получать информацию о наличии в кадре визем из заданного базового набора и проводить обучение правильной артикуляции для конкретного человека. Проведённые экспериментальные исследования показали возможность использования предложенной модели артикуляции для идентификации базовых состояний губ на тестовой выборке видеофрагментов 55 слов украинской речи.

• *мимика* • *визема* • *артикуляция* • *чтение по губам* • *жестовая речь*

An technology to the analysis of visemes of visual component of speech process in the video stream is proposed in this paper. The approach allows to compute information about presence of a viseme on an animation frame, choosing from a given base set to conduct tuition of correct articulation for a particular person. Experimental studies have shown the efficiency of using the mathematical model of lips presented in technology to identify the basic condition of lip articulation on test video samples with 55 words of the Ukrainian language.

• *mimics* • *visemes* • *articulation* • *lipsreading* • *sign language*

Синтез и анализ жестикуляции и выражений лица стали важной частью различных мультимедийных систем и информационных технологий интеллектуализации компьютерных интерфейсов, используя визуальную информацию. Такие систе-

мы особенно важны для людей, имеющих проблемы со слухом, ведь они компенсируют потерю звукового канала восприятия информации зрительным благодаря восприятию невербальной мимики, жестикуляции и чтения по губам [1, 2]. Даже люди с нормальным слухом и навыком речевого общения подсознательно используют информацию о движении губ и выражении лица, что было подтверждено в работе [3]. Исследования данной проблемы сориентированы на создание систем аудиовизуального распознавания речи, в которых используют визуальную информацию как дополнение к звуковой или ограничиваются только визуальной. В первом и втором направлении важной задачей будет определение базовых элементов визуального речевого потока или построение модели речи, дающей ответ на тот же вопрос: по последовательности изображений процесса артикуляции установить, что же было произнесено.

Хотя визуальный алфавит и является неполным, на практике он широко используется сурдопереводчиками жестового языка, дополняя жестикуляцию в тех случаях, когда необходимо дословно передать смысл предложения, информационного сообщения, сохранив грамматическую структуру предложения разговорной речи. Одним из перспективных направлений разработки систем обучения жестовому языку является создание системы обучения правильной артикуляции, основной задачей которой была бы возможность моделировать и контролировать правильность артикуляции губ при произношении слов некоторого языка, сравнивая её с эталонной. В качестве языка в настоящем исследовании будет использоваться украинский язык.

Визуальный алфавит

В процессе речеобразования при моделировании звуков соответствующих фонем на лице человека, вследствие движения мышц, отвечающих за артикуляцию, возникают различные мимические состояния. Фонемы, которые выглядят подобными друг другу при их артикуляции, можно отнести к одной группе, которая называется виземой [4]. Визема – характерное выражение лица, которое является визуальным портретом фонемы или иной базовой звуковой единицы в разговорной речи. Виземы есть те элементы, которые анализируются и распознаются в системах чтения по губам.

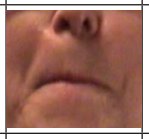
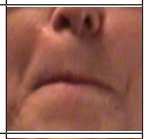
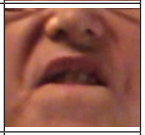
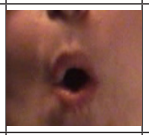
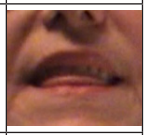
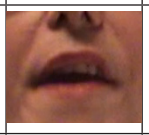
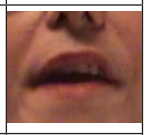
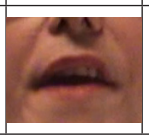
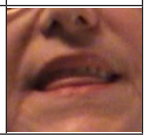
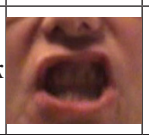
Визуальный алфавит может состоять из различного количества элементов для одного и того же языка в зависимости от уровня реалистичности восприятия артикуляции губ. В нём нет однозначного соответствия между произнесённой фонемой и её визуальным отображением, что снижает возможности зрительного восприятия речи. То есть довольно сложно по визуальному портрету фонемы восстановить её звуковой образ. Следует заметить, что фонетическая структура украинского языка, сама фонетика и артикуляция являются уникальными [5]. В украинском языке всего 38 фонем, из них 6 гласных и 32 согласных. Подробный анализ описания способа образования гласных и согласных [5, 6] позволил выделить для каждой фонемы те внешние артикуляторы, которые формируют соответствующую ей визему.

Все гласные звуки отражаются на лице различными положениями губ, языка, выражения лица, поэтому каждой фонеме соответствует собственная

визема. Поскольку выражения лица и положение губ при произношении звуков «ы» и «и» очень похожи, то для упрощения их часто рассматривают как одну визему. Таким образом, в украинском языке следует выделять 5 визем, соответствующих гласным фонемам («а», «о», «у», «е», «ыи»). Сложность определения характерных визуальных проявлений при артикуляции согласных языка объясняется активной ролью прежде всего языка или глотки в образовании щелей внутри речевого аппарата для формирования соответствующих звуков. Исключением являются губные согласные: фонемы «б», «м», «п», «ф» и «в» достаточно чётко визуально различаются.

Таблица 1

Виземы украинского языка

n	Визема	Фонемы		n	Визема	Фонемы	
1	«а»	а		9	«пбм»	п, б, м	
2	«е»	е		10	«вф»	в, ф	
3	«о»	о		11	«тднл»	т, д, н, л	
4	«у»	у		12	«сзцДз»	с, з, ц, дз	
5	«иы»	и, ы		13	«р»	р	
6	«й»	й		14	«л'р'»	л', р'	
7	«шжчДж»	ш, ж, ч, дж		15	«т'д'н'»	т', д', н'	
8	«кгхґ»	к, г, х, ґ		0	«спокойствие»	н/а	

В результате классификации фонем украинской речи получаем набор из 13(15) визем плюс состояние спокойствия (табл. 1). Отметим, что детализация визем для мягких фонем – л', р', т', д', н' – имеет большое значение при синтезе визуальной составляющей артикуляции.

Таким образом, специфика артикуляции украинского языка требует адаптации существующих и разработки новых методов анализа визуальной составляющей артикуляционного процесса [7]. При этом можно выделить следуючие задачи, которые необходимо решить:

1. Предварительная обработка в качестве этапа подготовки к выделению признаков состояния губ на изображении.
2. Определение множества характеристических признаков.
3. Выбор и применение методов класификации и кластеризации для определения, к какому классу следует отнести входную информацию о состоянии губ. Выбор математической модели губ будет определять способ получения численных характеристик процесса артикуляции и возможность его дальнейшего анализа.

Поэтому для синтеза математической модели предлагается перейти от набора фотографических изображений лица человека с процессом артикуляции к множеству векторов характеристических признаков, полученных из этих изображений. Процедуру такого перехода осуществим в несколько этапов: 1) на фото выделяются внутренние контуры губ; 2) полученные пиксельные значения внутренних контуров губ аппроксимируются с помощью неравномерных базисных сплайнов (NURBS); 3) на основании результатов NURBS-аппроксимации формируется вектор характеристических признаков.

Предварительная обработка и модель данных

Отметим, что визуальная речевая информация может быть искажена шумом и содержать большое количество несущественных деталей. В этом случае следует использовать методы и подходы для выделения полезной информации на изображениях, в частности, необходимо осуществить сегментацию изображений путём определения кромок и границ или локализацией объектов. Ниже приведён алгоритм выделения области губ на изображении.

Шаг 1. Выделение на изображении внутреннего контура губ:

$$Im\ L \rightarrow D, \tag{1}$$

где $Im\ L=\{L_k: I_k \in FSV\}$ – упорядоченное множество ключевых кадров видеопотока *FSV* (Face Speech Video), сформированного при съёмке мимических проявлений на лице человека, а именно положений губ, при проговаривании слов украинского языка (индекс k=1,N отвечает за порядковый номер кадра у выбранной последовательности, где N – количество ключевых кадров);

$I_k = \left\{col_{ij}^k\right\}_{i,j=1}^{m,n}, i = 1,\cdots,m; j = 1,\cdots,n$ – изображение размера mxn лица с мимическим положением губ, 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 кадра с номером k.

Шаг 2. Аппроксимация полученного внутреннего контура губ с помощью NURBS-кривых с целью получения вектора характеристических признаков:

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

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

Для выделения контуров губ разработан алгоритм с использованием адаптированного подхода выделения области губ на фотографическом изображении лица человека в процессе артикуляции слов на украинском языке. Таким образом уменьшается размерность входной информации для дальнейшей обработки изображения, отделяется область с губами от фона и других частей лица, что в дальнейшем облегчает поиск контура губ.

Общая схема алгоритма показана на рис. 1, где следует выделить два логических блока: «Блок предварительной обработки», который отвечает за преобразование входного изображения в специальное мозаичное представление, и «Блок комплексного поиска особых областей лица», результатом работы которого являются координаты и размеры зон лица, связанных с глазами и губами. В блоке предварительной обработки, для исключения влияния освещённости, был осуществлён переход к бинарному изображению, то есть переход изображения I к изображению в серых тонах, линейное и нелинейное выравнивания освещённости, сглаживание, выделение кромок. Для предварительной обработки визуальной видеоинформации использовалась упаковка EmguCV [8] библиотеки алгоритмов компьютерного зрения, обработки изображений и численных алгоритмов общего назначения с открытым кодом библиотеки OpenCV, которая имеет достаточно широкую функциональность для быстрой цифровой обработки видеоизображений.

Для нахождения областей использовался метод поиска по шаблону[9]. Используя базу с 28 фотографий лиц людей на основе их бинарного представления, экспериментально были получены шаблоны лица, рта и глаз. Для определения лучшего взаимного положения анализировался дискретный функционал качества $F_{optimal}$.

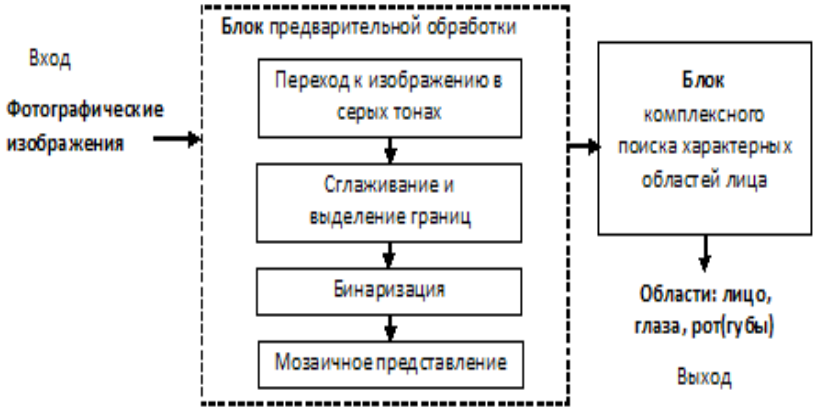


Рис.1. Общая схема алгоритма комплексного поиска областей лица человека на фотографическом изображении

$$F_{optimal}(i,j) =$$
$$= \alpha \cdot D_{face}(i,j) + \beta \cdot D_{eyes}(eyes(i)\ eyes(j)\ + \gamma \cdot D_{mouth}(mouth(i)\ mouth(j)),$$
$$F_{optimal}(i,j)=\alpha D_{face}(i,j)+\beta D_{eyes}(eyes(i),eyes(j)+\gamma D_{mouth}(mouth(i),mouth(j)), \quad (3)$$

где (i,j)– координаты пикселя матрицы изображения; $eyes(i)$, $eyes(j)$ – предполагаемые координаты положения области глаз, при условии, что левый верхний угол области лица соответствует координатам (i,j); D_{face}, D_{eyes}, D_{mouth} – отличие масок эталонов лица, глаз, рта от изображения; α, β, γ – весовые коэффициенты.

Экспериментально были установлены величины параметров: α=0.3, β=0.6, γ=0.1, из расчёта, что наиболее информативной считается область глаз. При этом

$$D_{face} = (\sum_{i,j}(\hat{I}(i,j) - E_{face}(i,j)\ ^2)^{1/2},$$

где $\hat{I}$ – определённым образом обработанное изображение I, E_{face} –шаблон лица. Положение областей определялось по формуле:

$$(i^*, j^*) = \arg\max_{i,j} F_{optimal}(i,j),$$

где (i^*, j^*) – координаты левого верхнего угла области лица, а $(eyes(i^*), eyes(j^*))$ и $(mouth(i^*), mouth(j^*))$ – глаз и рта соответственно.

Эффективными средствами для решения задач выделения контуров, сегментации, определения границ объекта являются модели, которые деформируются [10], их частным случаем могут быть NURBS-кривые [11]. Преимущества моделирования контуров лица или губ с помощью NURBS-кривых заключаются в следущем: 1) размерность уменьшается на порядки; 2) деформации кривых более плавные – подобные реальным мимическим проявлениям на лице человека.

В результате математической моделью мимических проявлений губ для речевых сигналов будет векторное пространство контрольных точек NURBS-кривых (рис. 2):

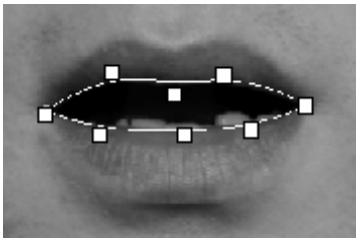


Рис. 2. Опорные точки NURBS-кривых для моделирования контуров губ

$$P = \{v : v = (x_0^{p_{top}}, \cdots, x_{n_{top}-1}^{p_{top}}, x_0^{p_{bot}}, \cdots, x_{n_{bot}-1}^{p_{bot}}, y_0^{p_{top}}, \cdots, y_{n_{top}-1}^{p_{top}}, y_0^{p_{bot}}, \cdots, y_{n_{bot}-1}^{p_{bot}})\}$$

$$p_j^{p_{[top|bot]}} = (x_j^{p_{[top|bot]}}, y_j^{p_{[top|bot]}}), \quad j = \overline{0, n_{[top|bot]} - 1},$$

где $v \epsilon P$ – вектор координат контрольных точек $P^{Pbot}{}_j$ и $P^{Ptop}{}_j$ апроксимирующих NURBS-кривые $Pbot^{(u)}$, $Ptop^{(u)}$, а n_{bot}, n_{top} – количество контрольных точек для NURBS-кривых $Pbot^{(u)}$, $Ptop^{(u)}$ соответственно. Размерность P определяется как M=2·(n_{top}+n_{bot}). В данном исследовании для постро-

ения базовых характеристических векторов рассматривается представление виземы одним кадром (изображением). Таким образом, для каждого класса фонем (см. табл. 1) получается его визема в виде кадра изображения, на основе которой строится вектор характеристических признаков (4).

Технология чтения по губам

Схема технологии чтения по губам (распознавания мимики) при произнесении слов на украинском языке показана на рис. 3.

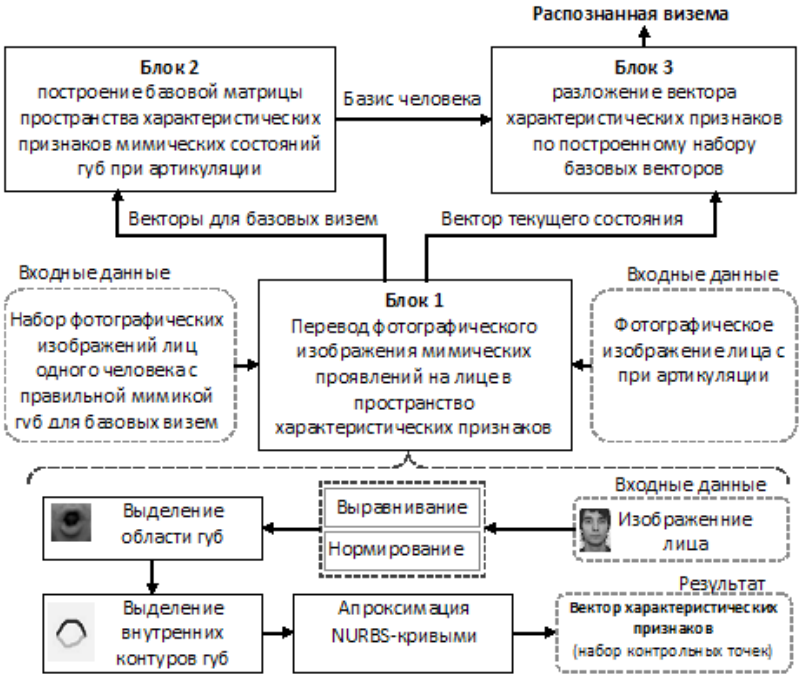


Рис.3 Схема чтения по губам слов на украинском языке

Блок 1 отвечает за предварительную обработку входящей визуальной информации и преобразование её во множество характеристических признаков (4).

Блок 2 включает в себя алгоритмы построения «базиса» (базовой матрицы) множества характеристических признаков и оценки его качества. На выходе строится базовая матрица A – матрица базовых состояний размера MxL характеристических признаков P. Она состоит из отнормированных и отцентрированных векторов, соответствующих базовым виземам: $A=(a_{ij})_{LxM}$, где $a_{ij} = v_j^{base_i} \in P$, L – количество базовых векторов, M– их размерность. Мимик или базовых визем, учитывая состояние покоя, для украинского языка удалось выделить шестнадцать.

В третьем блоке реализовано распознавание входных визем, путём проектирования вектора характеристических признаков b, построенного для входного изображения, на полученный базовый набор. В этом случае задача сводится к нахождению всех векторов, для которых выполняется:

$$Ax=b. \quad (5)$$

В случае, когда $\det(A^TA)>\varepsilon>0$, где T – знак транспонированния, решение задачи получается методом наименьших квадратов в виде выпуклой комбинации: $x=(A^TA)^{-1}A^Tb$. В противном случае наиболее надёжным методом для решения подобных задач является метод сингулярного разложения SVD [12].

Результатом работы технологии является вектор, на основе значений его компонентов принимается решение о соответствии входного вектора конкретным базовым состояниям (мимикам) при артикуляции слов на украинском языке.

Обучение правильной артикуляции

Логика работы технологии для решения проблемы обучения конкретного человека правильной артикуляции состоит в возможности получения базовых мимик. Алгоримически процесс обучения будет заключаться в следующем:

1. С помощью специалиста по дефектологии или профессионального сурдопереводчика (экспертов по разборчивости артикуляции губ) для данного человека устанавливается экспериментальным образом набор образцов визем языка (табл. 1), которые, по мнению экспертов, являются удовлетворительными (корректными) с точки зрения понятности (разборчивости) элементов артикуляционного процесса. Назовем это множество образцов «корректными образцами базовых визем», а её элементы «корректным образцом».

Следует отметить, что требование корректности для артикуляции слова человеком не может быть жёстко формализовано, учитывая существование присущих каждому человеку особенностей внешней артикуляции вследствие различных физиологических возможностей открытия и закрытия рта, формы губ и др. Корректность может быть представлена набором рекомендаций и правил артикуляции звуков, установленных для классического произношения (например, по трудам [5, 6]), и детальных описаний самих визем (пример для гласных фонем, табл. 2).

2. Для получения «корректных» базовых визем языка производится запись на видео процесса артикуляции человеком набора слов из обучающей выборки, покрывающей все возможные комбинации фонем. Каждое слово анализируется экспертом на предмет правильности (разборчивости) артикуляции.

3. В случае, когда находятся ошибки (то есть погрешность артикулирования слова является критической для того чтобы различать его образ), даются соответ- ствующие рекомендации по коррекции артикуляции данного слова и проводится перезапись слова до тех пор, пока не достигается приемлемый для эксперта результат.

Таким образом, получение выборки базовых визем не является неестественным и не- удобным для человека. Уникальность артикуляции каждого человека, учитывая присущие только ей особенности движения внешних артикуляторов, не является проблемным моментом вследствие того, что при формировании базового множества визем был достигнут необходимый уровень разборчивости.

4. По выборке базовых визем для человека строится базовая матрица, которая в дальнейшем используется для технологии чтения по губам.

Таблица 2

Детальные описания гласных фонем украинского языка

	Визема	Детальное описание
«а»		Губы в создании активного участия не принимают; образуют большое отверстие и прижаты к зубам: не выпячиваются и не растянуты в стороны; тело языка при этом заметно оттянуто назад – к стенке глотки.
«э»		Ротовая полость открыта больше, чем при создании «**и**» и «**ы**», приближаясь в какой-то мере к громкому «**а**»; уголки губ при произношении «**э**» немного растянуты в стороны и прижаты к зубам, положение языка при произношении «**э**» в целом среднее и при этом значительно ниже, чем при произношении гласных «**ы**» и «**и**».
«ы»		Ротовая полость более закрыта, чем при произношении «**э**»; уголки губ растянуты в стороны чуть больше, чем при произношении «**э**», и прижаты к зубам; передняя часть спинки языка выпуклая, основной своей массой продвинутая вперёд; положение языка выше, чем при произношении «**э**», но ниже, чем при произношении «**и**».
«и»		Ротовая полость раскрыта меньше; язык собран спереди, имеет отчётливо выпуклую форму, степень поднятия спинки языка к твёрдому нёбу при создании «и» наивысшая среди других гласных переднего ряда; губы, образуя узкую щель, растянуты в стороны чуть больше, чем при артикуляции «**ы**» и «**э**».
«о»		Ротовая полость открыта меньше, чем при артикуляции «**а**», и более, чем при «**у**»; губы при произношении «**о**» вытянуты вперёд, между ними образуется округлое отверстие; задняя часть спинки языка отодвинута назад.
«у»		Ротовая полость открыта меньше, чем при произношении «**о**» и «**а**», губы при артикуляции «**у**» отходят от передних зубов и очень вытягиваются вперёд, образуя между губами и зубами небольшую полость; язык направлен вверх к мягкому или даже к средней части твёрдого нёба.

Экспериментальные результаты и обсуждение

Для проверки эффективности предложенной технологии было реализовано соответствующее программное приложение (см. рис. 4) со следующей функциональностью:

1) выделение области лица из кадра изображения;
2) выделение контура губ и локализация характеристических точек;
3) определение вектора характеристических параметров;
4) распознавание конкретного статического состояния губ в кадре видеоизображения по построенной базовой матрице.

Для корректной работы программы на изображение или кадр видео накладываются следующие ограничения:

- лица человека на изображении должно занимать не менее 30% площади фотографии или кадра;
- лицо человека наклонено не более, чем под углом 15º, чтобы уголки губ были по вертикальной оси ниже, чем точки дуги Купидона;
- лицо человека должно быть освещено достаточно равномерно, и цвет губ должен существенно отличаться от цвета кожи;
- первый кадр видео соответствует состоянию покоя губ.

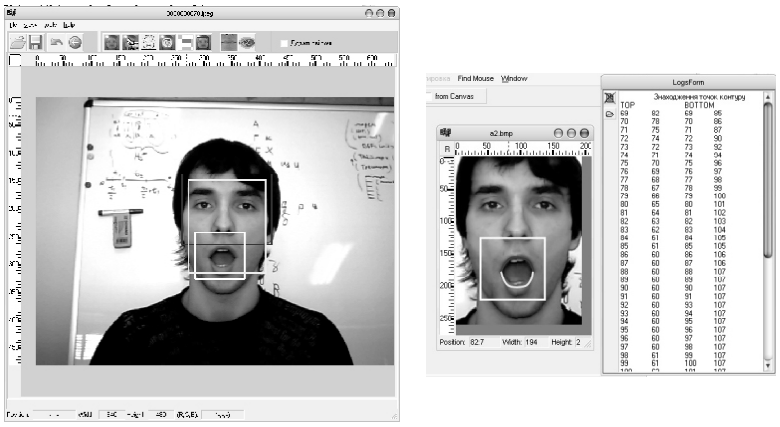


Рис. 4. Окна программного комплекса анализа артикуляционного процесса

Результаты правильного распознавания визем составляют около 92% на тестовой виборке с 20 кадров-визем одного человека, полученных при записи 55 слов украинской речи. В эти слова в фонетическом представлении входили звуки, которые соответствуют базовым виземам.

Выводы

Разработана технология для чтения по губам на основе гибких шаблонов, представленных с помощью неоднородных рациональных B-сплайнов с применением для решения задачи обучения правильной артикуляции.

В качестве характеристических признаков предлагается использовать параметры математической модели состояний губ при артикуляции. Для синтеза математической модели и поиска её параметров делается переход от фотографических изображений лица человека с процессом произнесения к множеству векторов характеристических признаков.

Дальнейшие исследования направлены на усовершенствования технологии за счёт учёта динамики изменения состояния губ, используя информацию из предыдущих кадров.

Литература

1. *Ouni S., Cohen M., Ishak H., Massaro D.* Visual contribution to speech perception: measuring the intelligibility of animated talking heads // Journal on Audio, Speech and Music Processing. — 2007. Issue 1. — P. 1–12.
2. *Воскресенский А.Л., Хахалин Г.К.* От звучащей речи – к жестовой // Речевые технологии. — 2009. — №1. — С. 99–106.
3. *McGurk H., MacDonald J.* Hearing lips and seeing voices // Nature. — 1976. — Vol. 264. — P.746–768.
4. *Fisher C.G.* Confusion among visually perceived consonants // Journal of Speech and Hearing Research. — 1968. Vol. 11. — P.796–804.
5. *Билодид И.К.* Современный украинский литературный язык. – К.: Наук. думка, 1969. — 435 с. (на украинском языке).
6. *Крак Ю.В., Бармак О.В., Тернов А.С.* Информационная технология для автоматического чтения по губам украинской речи // Комп'ютерна математика. — 2009. — № 1. — С. 86.
7. *Кривонос Ю.Г., Крак Ю.В., Тернов А.С.* Локализация и учёт особенностей лица человека для задач распознавания по портретной фотографии // Искусственный интеллект. — 2007. — № 3.
8. Электронный ресурс EmguCV. Режим доступа: http://www.emgu.com /wiki/index.php/Main_Page.
9. *Stan Z. Li, Jain Anil K.* Handbook of face recognition. Springer-Verlag London Limited. — 2005. — 395 p.
10. *Крак Ю.В. Бармак А.В., Ефимов Г.Н.* Использование контурных моделей для построения базиса пространства мимических выражений эмоций // Искусственный интеллект. — 2007. — № 4. — С. 288–296 (на украинском языке).
11. *Форсайт Дж.* Машинные методы математических вычислений. Пер. с англ. Х.Д Икрамова. — М.: Мир, 1980.— 277 с.
12. *Крак Ю.В. Тернов А.С., Лисняк М.П.* Структурно-виземный анализ артикуляции украинской речи // Искусственный интеллект. — 2011. — № 3. — С.156–166 (на украинском языке).

Сведения об авторах:

Крак Юрий Васильевич,
доктор физико-математических наук, профессор Киевского национального университета имени Тараса Шевченко, старший научный сотрудник института кибернетики им. В.М. Глушкова НАН Украины. Специалист в области искусственного интеллекта, анализа и синтеза голосовой и жестовой коммуникационной информации.

Тернов Антон Сергеевич,
кандидат технических наук, научный сотрудник института кибернетики им. В.М. Глушкова НАН Украины. Специалист в области искусственного интеллекта. Круг научных интересов включает распознавание образов, обработку и анализ изображений, виртуальную реальность, моделирование и распознавание жестовой речи.