УДК 20.19.29

ДВУЯЗЫЧНАЯ МНОГОМОДАЛЬНАЯ СИСТЕМА ДЛЯ АУДИОВИЗУАЛЬНОГО СИНТЕЗА РЕЧИ И ЖЕСТОВОГО ЯЗЫКА ПО ТЕКСТУ

А.А. Карпов[a], М. Железны[b]

[a] Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, 199178, Россия
[b] Западночешский Университет, Пльзень, 30614, Чехия, zelezny@kky.zcu.cz

Аннотация. Представлена концептуальная модель, архитектура и программная реализация многомодальной системы аудиовизуального синтеза речи и жестового языка по входному тексту. Основными компонентами разработанной многомодальной системы синтеза (жестовый аватар) являются: текстовый процессор анализа входного текста; имитационная трехмерная модель головы человека; компьютерный синтезатор звучащей речи; система синтеза аудиовизуальной речи; имитационная модель верхней части тела и рук человека; многомодальный пользовательский интерфейс, интегрирующий компоненты генерации звучащей, визуальной и жестовой речи по тексту. Предложенная система выполняет автоматическое преобразование входной текстовой информации в речевую (аудиоинформацию) и жестовую (видеоинформацию), объединение и вывод ее в виде мультимедийной информации. На вход системы подается произвольный грамматически корректный текст на русском или чешском языке, который анализируется текстовым процессором для выделения предложений, слов и букв. Далее полученная текстовая информация преобразуется в символы жестовой нотации (используется международная «Гамбургская система нотации» – HamNoSys, которая описывает основные дифференциальные признаки каждого жеста рук: форму кисти, ориентацию руки, место и характер движения), на основе которых трехмерный жестовый аватар воспроизводит элементы жестового языка. Виртуальная трехмерная модель головы и верхней части тела человека реализована на языке моделирования виртуальной реальности VRML и управляется программно средствами графической библиотеки OpenGL. Предложенная многомодальная система синтеза является универсальной, она предназначена как для обычных пользователей, так и для людей с ограниченными возможностями здоровья (в частности, глухих и незрячих людей) и служит для целей мультимедийного аудиовизуального вывода вводимой текстовой информации.

Ключевые слова: многомодальные интерфейсы пользователя, человеко-машинное взаимодействие, жестовый язык, синтез речи, трехмерные модели, ассистивные технологии, жестовый аватар.

Благодарности. Исследование выполнено при частичной финансовой поддержке Правительства Российской Федерации (грант № 074-U01), фонда РФФИ (проект № 12-08-01265_а) и Европейского фонда регионального развития (ЕФРР), проект «Новые технологии для информационного общества» (NTIS), Европейский центр передового опыта, ED1.1.00/02.0090.

BILINGUAL MULTIMODAL SYSTEM FOR TEXT-TO-AUDIOVISUAL SPEECH AND SIGN LANGUAGE SYNTHESIS

A.A. Karpov[a], M. Zelezny [b]

[a] Saint Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Saint Petersburg, 199178, Russia
[b] University of West Bohemia, Pilsen, 30614, Czech Republic, zelezny@kky.zcu.cz

Abstract. We present a conceptual model, architecture and software of a multimodal system for audio-visual speech and sign language synthesis by the input text. The main components of the developed multimodal synthesis system (signing avatar) are: automatic text processor for input text analysis; simulation 3D model of human's head; computer text-to-speech synthesizer; a system for audio-visual speech synthesis; simulation 3D model of human's hands and upper body; multimodal user interface integrating all the components for generation of audio, visual and signed speech. The proposed system performs automatic translation of input textual information into speech (audio information) and gestures (video information), information fusion and its output in the form of multimedia information. A user can input any grammatically correct text in Russian or Czech languages to the system; it is analyzed by the text processor to detect sentences, words and characters. Then this textual information is converted into symbols of the sign language notation. We apply international «Hamburg Notation System» - HamNoSys, which describes the main differential features of each manual sign: hand shape, hand orientation, place and type of movement. On their basis the 3D signing avatar displays the elements of the sign language. The virtual 3D model of human's head and upper body has been created using VRML virtual reality modeling language, and it is controlled by the software based on OpenGL graphical library. The developed multimodal synthesis system is a universal one since it is oriented for both regular users and disabled people (in particular, for the hard-of-hearing and visually impaired), and it serves for multimedia output (by audio and visual modalities) of input textual information.

Keywords: multimodal user interfaces, human-computer interaction, sign language, speech synthesis, 3D models, assistive technologies, signing avatar.

Acknowledgements. The research is partially financially supported by the Government of the Russian Federation (grant № 074-U01), by the Russian Foundation for Basic Research (project № 12-08-01265_a) and European Foundation of Regional Progress, project "New Technologies for Information Society" (NTIS), European Centre of Advanced Experience, ED1.1.00/02.0090.

Введение

Основным способом межчеловеческой коммуникации в обществе глухих и слабослышащих людей является жестовый язык (ЖЯ), в котором каждому смысловому понятию (или группе синонимичных понятий) соответствует определенный уникальный жестовый эквивалент. В нем для передачи информации используются различные визуально-кинетические средства естественного межчеловеческого общения: жесты рук, мимика и эмоции лица, артикуляция губ. ЖЯ не является универсальным во всех странах мира, так как он возникает и развивается естественным путем в разных локальных сообществах и изменяется со временем с появлением новой лексики. По данным международного интернет-каталога Ethnologue (www.ethnologue.com/subgroups/deaf-sign-language), в разных странах мира насчитывается порядка 140 различных жестовых языков, не считая их региолекты.

По статистике Министерства здравоохранения Российской Федерации (РФ), в России насчитывается около 200 тысяч глухих и слабослышащих граждан, имеющих инвалидность по слуху, и около 95% инвалидов по слуху общаются посредством языка жестов. Без его использования не обойтись также на приеме у врача, у нотариуса, на судебном процессе, на официальных лекциях, при бытовых разговорах и т.д. Также, согласно последней переписи населения 2010 г., в России насчитывается более 120 тысяч человек, владеющих русским ЖЯ. Для сравнения, в Чехии по статистике насчитывается около 7500 носителей чешского ЖЯ (это около 0,07% населения) и около 500 тысяч слабослышащих людей (4,7% всего населения страны) [1].

При этом нужно отметить, что ЖЯ является вторым государственным языком устного общения граждан в США, Финляндии, Испании, Чехии и т.д., что закреплено Конституцией этих стран. До недавнего времени русский ЖЯ не имел в России никакого официального статуса, однако в самом конце 2012 г. Президент РФ подписал закон, определяющий официальный статус русского ЖЯ в России как языка общения при наличии у людей нарушений слуха или речи, в том числе в сферах устного использования государственного языка РФ. В результате вступления в силу данного закона глухие люди получают возможность, например, обращения в государственные учреждения на ЖЯ. В дальнейшем предполагается создание автоматизированных систем субтитрирования и сурдоперевода телевизионных программ и кинофильмов.

Основной задачей автоматизированных систем, основанных на компьютерной обработке ЖЯ и речи, является обеспечение равноправной коммуникации слышащих и глухих людей с нарушениями слуха, которых во всем мире насчитываются десятки миллионов человек. Одним из самых эффективных средств обучения и взаимодействия являются мультимедийные компьютерные системы, поэтому создание информационных приложений, способных работать с ЖЯ, является одной из приоритетных задач при работе с глухими и слабослышащими людьми. Особый интерес в этой области представляют системы компьютерного синтеза жестового языка и речи по входному тексту.

Одним из наиболее эффективных вариантов для реализации компьютерных систем синтеза ЖЯ является использование трехмерных анимированных моделей человека (так называемых жестовых аватаров (signing avatar)), которые могут управляться посредством символов жестовой нотации, описывая требуемые конфигурации рук и различные типы движений. Жесты из лексикона в такой системе синтеза представляют собой цепочку символов в выбранной нотации, поэтому словарь может легко модифицироваться и расширяться без использования специального оборудования. Специфика автоматических систем человеко-машинного взаимодействия и коммуникации состоит в том, что ЖЯ и жестовый словарь должны быть определенным образом формализованы, чтобы компьютер мог обрабатывать и синтезировать жесты. Для описания жеста по его визуальным признакам существуют несколько различных систем нотации (например, HamNoSys или Sign Writing), позволяющих зафиксировать описание жеста. Довольно широкое распространение в последнее время в мире (особенно в странах Европы) получила «Гамбургская система нотации» (HamNoSys) [2, 3]. Эта система отличается наибольшей проработанностью инвентаря знаков и пригодна для использования в компьютерных приложениях за счет того, что ее знаки переведены в компьютерную систему кодировки Юникод с соответствующими компьютерными шрифтами. Инвентарь HamNoSys позволяет записать практически любой жест, выполняемый одной или двумя руками, что делает эту систему универсальной и пригодной для записи практически любого ЖЯ мира. Таким образом, наиболее перспективным вариантом использования трехмерных виртуальных аватаров для синтеза ЖЯ является их управление посредством символов нотации жестов, которые описывают требуемые конфигурации рук и различные типы движений. Словарь жестов в такой системе представляет собой цепочку символов в одной системе нотации, поэтому может легко модифицироваться и расширяться без использования специального оборудования.

Используя такой подход, за последние годы на волне создания ассистивных технологий был разработан ряд моделей компьютерного синтеза жестовой речи для нескольких ЖЯ, включая американский, британский, французский, чешский ЖЯ и др. Среди известных зарубежных компьютерных систем синтеза жестовой речи, использующих различные анимированные аватары, следуют отметить системы, разработанные в рамках различных проектов в Евросоюзе и США: DePaul ASL Synthesizer

(http://asl.cs.depaul.edu), аватары европейских проектов Dicta-Sign (www.dictasign.eu) [4], SIGNSPEAK (www.signspeak.eu/en) [5], SignCom [6], Italian SL [7], ViSiCAST (аватар Visia, www.visicast.co.uk), eSign (аватар vGuido, www.sign-lang.uni-hamburg.de/esign), аватары Sign Smith и Sign4Me компании Vcom3D (www.vcom3d.com), SiSi от IBM (www-03.ibm.com/press/us/en/pressrelease/22316.wss), американскую систему iCommunicator (www.icommunicator.com) и ряд других.

Модель и архитектура многомодальной системы синтеза речи и жестового языка

В ходе исследований авторами была разработана модель универсального многомодального человеко-машинного интерфейса (для вывода мультимедийной информации) и архитектура многомодальной системы синтеза аудиовизуальной речи и ЖЯ по тексту. Концептуальная модель универсального многомодального интерфейса пользователя представлена на рис. 1. Интерфейс выполняет автоматическое преобразование входной текстовой информации T в жестовую G, а также речевую аудиоинформацию A и видеоинформацию V, объединение и вывод ее в виде мультимедийной информации M (при этом преобразование g является функцией автоматической обработки текста, а f выполняет функцию объединения разнотипной информации в мультимедийное представление):

$$T \xrightarrow{g} < G, A, V > \xrightarrow{f} M .$$

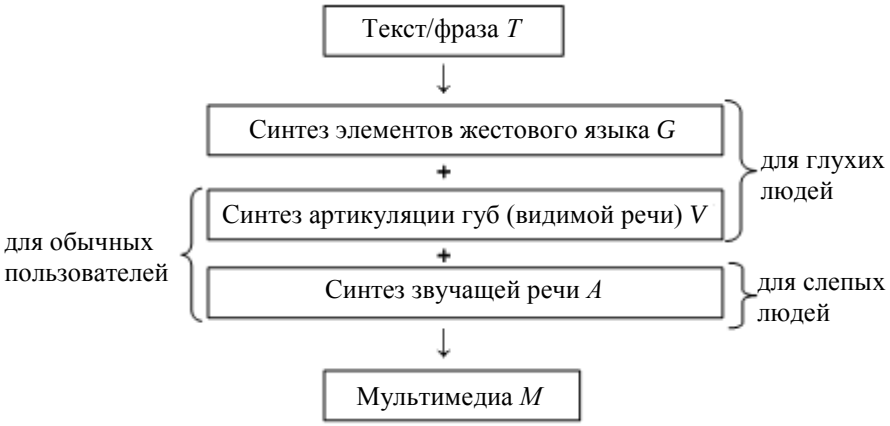

Рис. 1. Концептуальная модель универсального многомодального человеко-машинного интерфейса на основе синтеза аудиовизуальной речи и жестового языка

Такой пользовательский интерфейс вывода информации является универсальным, так как он предназначен для вывода входных текстовых данных посредством синтеза звучащей речи, артикуляции губ аватара и жестового языка как для обычных пользователей, так и для людей с ограниченными возможностями (глухих и незрячих людей).

На рис. 2 представлена разработанная архитектура многомодальной системы синтеза аудиовизуальной речи и ЖЯ по тексту [8]; ее основными компонентами являются:

- текстовый процессор анализа входного текста для последующего аудиосинтеза звучащей речи (по словам) и видеосинтеза жестовой и дактильной речи (показ фраз по словам или по буквам);
- имитационная трехмерная модель головы человека [9];
- аудиосинтезатор звучащей речи, осуществляющий преобразование текст–речь по входному тексту [10, 11];
- компьютерная система синтеза аудиовизуальной речи (говорящая голова (talking head)) на основе виртуальной объемной модели головы человека и машинного синтеза речи [12, 13];
- имитационная трехмерная модель верхней части тела и рук человека, в которой настраиваются параметры движений рук для синтеза элементов жестового языка на основе управляющих символов жестовой нотации HamNoSys [2, 14];
- многомодальный пользовательский интерфейс, интегрирующий компоненты генерации звучащей, визуальной и жестовой речи по входному тексту [1].

В многомодальной системе артикуляция губ, которые являются видимой частью органов речеобразования, сопровождается также синтезированной речью, которая может и не быть доступна полностью глухим людям, однако для слышащих людей синтез аудиовизуальной речи доступен и даже необходим для повышения разборчивости и естественности синтезируемой компьютером речи.

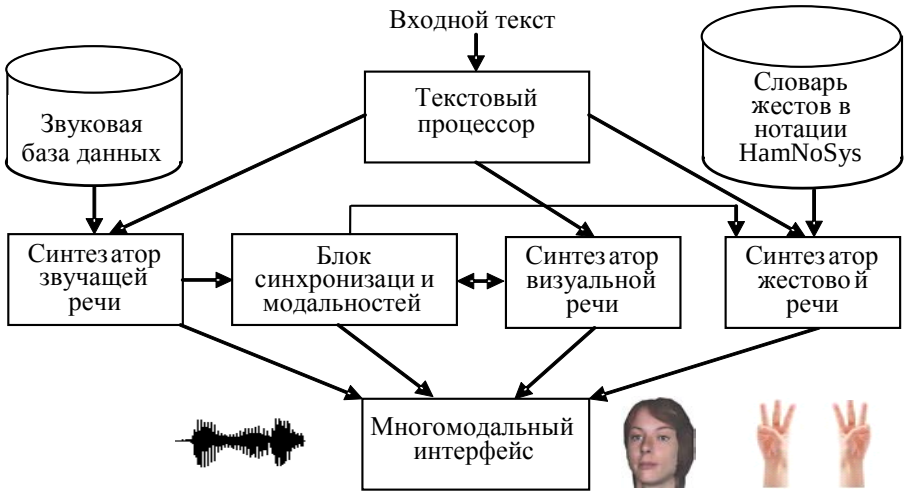

Рис. 2. Архитектура многомодальной системы синтеза аудиовизуальной речи и жестового языка по тексту

Программная реализация двуязычной многомодальной системы синтеза

В качестве практической реализации универсального многомодального интерфейса была программно разработана двуязычная многомодальная система синтеза аудиовизуальной речи и элементов ЖЯ. За основу была взята компьютерная система видеосинтеза чешского ЖЯ [15], созданная ранее Западночешским университетом в рамках исследовательского проекта Musslap (http://musslap.zcu.cz), которая была адаптирована к обработке русского ЖЯ. В итоге была разработана не просто система видеосинтеза жестов рук трехмерного аватара, а многомодальная система синтеза, в которой, помимо видеообработки, выполняется также аудиовизуальный синтез звучащей речи (хотя для глухих людей она и не доступна, но необходима для правильного синтеза артикуляции губ и мимики лица).

На вход системы подается произвольный грамматически корректный текст (на русском или чешском языке), который анализируется текстовым процессором, в нем выделяются предложения, слова (для аудиосинтеза речи и видеосинтеза артикуляции губ аватара) и буквы (для машинного синтеза дактильной речи [16]), которые автоматически преобразуются в символы жестовой нотации, на основе которой аватар воспроизводит мануальные жесты, декодируя символы нотации. Элементы жестовой речи в системе описываются при помощи системы записи жестов HamNoSys, отражающей основные дифференциальные признаки каждого жеста: форму кисти, ориентацию руки, место и характер движения. Виртуальная модель головы и верхней части туловища человека реализована на языке моделирования виртуальной реальности (Virtual Reality Modelling Language, VRML) и управляется программно средствами графической библиотеки OpenGL под управлением операционной системы семейства Microsoft Windows. Создан трехмерный жестовый аватар (signing avatar), который может иметь светлую либо черную одежду (в зависимости от предпочтений пользователя). Данный аватар, демонстрирующий финальную статью динамического жеста для числительного «16», показан в различных проекциях на рис. 3.

Рис. 3. Трехмерный жестовый аватар в различных проекциях

Подсистема аудиовизуального синтеза речи (так называемая говорящая голова [9, 17]) производит компиляционный синтез звучащей разговорной речи по тексту, совмещенный с движениями губ и лицевых органов трехмерной виртуальной головы человека. Для синтеза ЖЯ визуальная модель головы человека объединена с моделью туловища и рук (включая пальцы) человека. Как уже говорилось, ЖЯ складывается из комбинации динамических жестов, воспроизводимых обеими руками (либо одной рукой), и артикуляции губ, проговаривающих показываемое слово. При этом многие глухие люди, которые в детстве обладали слухом, способны «читать речь по губам» собеседника даже без использования жестов руками, поэтому такая речевая модальность должна являться неотъемлемым элементом компьютерной системы синтеза жестовой речи.

Также для системы многомодального синтеза разработан специальный метод синхронизации выходных аудио- и видеомодальностей. Синхронизация звучащей речи и жестов в системе осуществляется на основе временных меток начала и конца слов звучащей речи, синтезированной системой по тексту. Так как звучащая речь в среднем имеет более высокий темп воспроизведения, чем жестовая речь, то виртуальный аватар последовательно проговаривает и артикулирует с невысокой скоростью изолированные слова звучащей речи, дожидаясь окончания жестикуляции соответствующего слова (может включать в себя несколько последовательных букв дактильной азбуки), плавно переходя к следующему жесту слитной жестовой речи.

Разработанный жестовый аватар максимально имитирует стиль жестикуляции живых людей. Так, все жесты следуют в речи без пауз с соблюдением «плавности» и «текучести» жестов, что позволяет оформлять целые фразы и лексемы, а не набор изолированных друг от друга жестов. На данный момент словарь жестов системы составляет несколько сотен жестов для наиболее распространенных слов, цифр, букв и т.д.

Следует отметить, что компьютерный синтез ЖЯ с использованием трехмерного аватара обладает рядом достоинств при организации вывода информации пользователям с нарушениями слуха, в частности:

1. позволяет просматривать видеосинтез жестовой речи с разных сторон и углов обзора, что дает возможность лучше воспринимать пространственную информацию, например, степень удаленности рук от туловища и друг относительно друга (в отличие от двухмерных моделей);
2. дает возможность относительно легко пополнять и корректировать словарь жестов, так как вместо видеозаписей реального человека в словаре присутствуют компьютерные анимированные аватары, соответственно, для расширения словаря не обязательно записывать того же самого человека-демонстратора жестов в той же самой одежде и с той же прической, а также уровнем освещенности;
3. позволяет выполнять слитный синтез ЖЯ, в котором отдельные слова во фразах стыкуются бесшовно, т.е. не видны явные границы между соседними словами;
4. дает возможность заменять виртуальный аватар, используя новые высокореалистичные модели людей (мужчин или женщин, а также любых персонажей);
5. позволяет воспроизводить жестовую речь на экране с любой необходимой скоростью, как замедляя, так и ускоряя видеоряд.

Демонстрация и тестирование системы в Санкт-Петербурге были организованы при помощи сотрудников «Всероссийского общества глухих». Отзывы и качественная оценка системы потенциальными пользователями позволяют говорить об обеспечении естественности и разборчивости синтезированных элементов русского ЖЯ и дактильной речи (дактилологии), а также артикуляции и мимики губ виртуального аватара при речеобразовании.

Разработанная многомодальная система аудиовизуального синтеза речи и жестового языка (жестовый аватар) предназначена для организации универсальных человеко-машинных интерфейсов с целью коммуникации с людьми, имеющими тяжелые нарушения слуха и полностью глухими, посредством элементов разговорного ЖЯ (калькирующей жестовой речи и дактильной речи, воспроизводимых движениями/жестами рук виртуального помощника–аватара) и визуальной речи (артикуляции губ, обязательно сопутствующей жестовой модальности), а также речевой коммуникации со слепыми и слабовидящими людьми и полноценного мультимедийного общения со зрячими и слышащими пользователями. Данная система может использоваться для задач организации коммуникации человек–человек и человек–машина [1], в системах электронного обучения [18, 19], машинного перевода [20], виртуальной и дополненной реальности и т.д.

Представленная многомодальная система аудиовизуального синтеза речи и ЖЯ входит в состав большего комплекса – универсальной ассистивной информационной технологии (assistive technology) для людей с ограниченными возможностями здоровья [21], в состав которой также входят многомодальная система для бесконтактной работы с компьютером [22, 23], система аудиовизуального распознавания речи [24], система автоматического распознавания элементов жестового языка [25] и модель ассистивного интеллектуального пространства [26].

Заключение

В работе был представлен универсальный многомодальный интерфейс и разработана программная компьютерная система для аудиовизуального синтеза элементов жестового языка и звучащей речи по тексту, объединяющая бимодальную систему синтеза речи (виртуальная говорящая голова), обеспечивающую аудиовизуальный синтез речи, и модель тела и рук человека (жестовый аватар), выполняющую видеосинтез динамических жестов. Система предназначена для вывода входных текстовых данных посредством синтеза звучащей речи, артикуляции губ аватара и жестового языка как для обычных пользователей, так и для людей с ограниченными возможностями (глухих и незрячих).

Следующим этапом научно-исследовательских работ будет являться разработка системы синтеза разговорного жестового языка и речи по тексту. Ее создание осложняется необходимостью средств машинного перевода текста на жестовый язык, обладающий собственной структурой и грамматикой (отличной от письменного или устного языка), которые пока слабо исследованы лингвистами и недостаточно формализованы, но исследования в данном направлении активно ведутся, что позволяет говорить о скором решении данной проблемы. Дальнейшие исследования, разработки и внедрение в жизнь глухих людей автоматизированных компьютерных систем должны дополнительно привлечь к этой проблеме внимание общественности, а также обратить усилия ученых, разработчиков и компьютерных лингвистов к междисциплинарным исследованиям и работам в этой сфере обработки естественного языка.

References

1. Karpov A., Krnoul Z., Zelezny M., Ronzhin A. Multimodal synthesizer for Russian and Czech sign languages and audio-visual speech. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8009 LNCS, part 1, pp. 520–529. doi: 10.1007/978-3-642-39188-0-56
2. Hanke T. HamNoSys – representing sign language data in language resources and language processing contexts. *Proc. International Conference on Language Resources and Evaluation, LREC 2004*. Lisbon, Portugal, 2004, pp. 1–6.
3. Karpov A.A., Kagirov I.A. Formalizatsiya leksikona sistemy komp'yuternogo sinteza yazyka zhestov [Lexicon formalization for a computer system of sign language synthesis]. *SPIIRAS Proceedings*, 2011, no. 1 (16), pp. 123–140.
4. Efthimiou E. et al. Sign language technologies and resources of the dicta-sign project. *Proc. 5th Workshop on the Representation and Processing of Sign Languages*. Istanbul, Turkey, 2012, pp. 37–44.
5. Caminero J., Rodríguez-Gancedo M., Hernández-Trapote A., López-Mencía B. SIGNSPEAK project tools: a way to improve the communication bridge between signer and hearing communities. *Proc. 5th Workshop on the Representation and Processing of Sign Languages*. Istanbul, Turkey, 2012, pp. 1–6.
6. Gibet S., Courty N., Duarte K., Naour T. The SignCom system for data-driven animation of interactive virtual signers: methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems*, 2011, vol. 1, no. 1, art. 6. doi: 10.1145/2030365.2030371
7. Borgotallo R., Marino C., Piccolo E., Prinetto P., Tiotto G., Rossini M. A multi-language database for supporting sign language translation and synthesis. *Proc. 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. Malta, 2010, pp. 23–26.
8. Karpov A.A. Komp'yuternyi analiz i sintez russkogo zhestovogo yazyka [Computer analysis and synthesis of Russian sign language]. *Voprosy Yazykoznaniya*, 2011, no. 6, pp. 41–53.
9. Železný M., Krňoul Z., Císař P., Matoušek J. Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis. *Signal Processing*, 2006, vol. 86, no. 12, pp. 3657–3673. doi: 10.1016/j.sigpro.2006.02.039
10. Tihelka D., Kala J., Matoušek J. Enhancements of viterbi search for fast unit selection synthesis. *Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH-2010*. Makuhari, Japan, 2010, pp. 174–177.
11. Hoffmann R., Jokisch O., Lobanov B., Tsirulnik L., Shpilewsky E., Piurkowska B., Ronzhin A., Karpov A. Slavonic TTS and SST conversion for let's fly dialogue system. *Proc. 12th International Conference on Speech and Computer SPECOM-2007*. Moscow, Russia, 2007, pp. 729–733.
12. Krňoul Z., Železný M., Müller L. Training of coarticulation models using dominance functions and visual unit selection methods for audio-visual speech synthesis. *Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH*. Pittsburgh, USA, 2006, vol. 2, pp. 585–588.
13. Karpov A., Tsirulnik L., Krňoul Z., Ronzhin A., Lobanov B., Železný M. Audio-visual speech asynchrony modeling in a talking head. *Proc. Annual Conference of the International Speech Communication Association INTERSPEECH*. Brighton, UK, 2009, pp. 2911–2914.
14. Krňoul Z., Železný M. Translation and conversion for Czech sign speech synthesis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007, pp. 524–531.

15. Krňoul Z., Kanis J., Železný M., Müller L. Czech text-to-sign speech synthesizer. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 4892 LNCS, pp. 180–191. doi: 10.1007/978-3-540-78155-4_16
16. Karpov A.A. Mashinnyi sintez russkoi daktil'noi rechi po tekstu [Computer synthesis Russian finger spelling by text]. *Nauchno-Tekhnicheskaya Informatsiya. Seriya 2: Informatsionnye Protsessy i Sistemy*, 2013, no. 1, pp. 20–26.
17. Karpov A. A., Tsirulnik L. I., Zelezny M. Razrabotka komp'yuternoi sistemy “govoryashchaya golova” dlya audiovizual'nogo sinteza russkoi rechi po tekstu [Development of a computer system "Talking Head" for text-to-audiovisual-speech synthesis]. *Informatsionnye Tekhnologii*, 2010, no. 8, pp. 13–18.
18. Borgia F., Bianchini C.S., De Marsico M. Towards improving the e-learning experience for deaf students: e-LUX. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8514 LNCS, part 2, pp. 221–232. doi: 10.1007/978-3-319-07440-5_21
19. Tampel I.B., Krasnova E.V., Panova E.A., Levin K.E., Petrova O.S. Ispol'zovanie informatsionno-kommunikatsionnykh tekhnologii v elektronnom obuchenii inostrannym yazykam [Application of information and communication technologies in computer aided language learning]. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2013, no. 2 (84), pp. 154–160.
20. Hruz M., Campr P., Dikici E., Kindiroğlu A.A., Krňoul Z., Ronzhin A., Sak H., Schorno D., Yalçin H., Akarun L., Aran O., Karpov A., Saraçlar M., Železný M. Automatic fingersign to speech translation system. *Journal on Multimodal User Interfaces*, 2011, vol. 4, no. 2, pp. 61–79. doi: 10.1007/s12193-011-0059-3
21. Karpov A., Ronzhin A. A universal assistive technology with multimodal input and multimedia output interfaces. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8513 LNCS, part 1, pp. 369–378. doi: 10.1007/978-3-319-07437-5_35
22. Karpov A.A. ICanDo: Intellektual'nyi pomoshchnik dlya pol'zovatelei s ogranichennymi fizicheskimi vozmozhnostyami [ICanDo: Intelligent assistant for users with physical disabilities]. *Vestnik Komp'yuternykh i Informatsionnykh Tekhnologii*, 2007, no. 7, pp. 32–41.
23. Karpov A., Ronzhin A., Kipyatkova I. An assistive bi-modal user interface integrating multi-channel speech recognition and computer vision. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6762, part 2, pp. 454–463. doi: 10.1007/978-3-642-21605-3_50
24. Karpov A., Ronzhin A., Markov K., Zelezny M. Viseme-dependent weight optimization for CHMM-based audio-visual speech recognition. *Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. Makuhari, Japan, 2010, pp. 2678–2681.
25. Kindiroglu A., Yalcın H., Aran O., Hruz M., Campr P., Akarun L., Karpov A. Automatic recognition of fingerspelling gestures in multiple languages for a communication interface for the disabled. *Pattern Recognition and Image Analysis*, 2012, vol. 22, no. 4, pp. 527–536. doi: 10.1134/S1054661812040086
26. Karpov A.A., Akarun L., Ronzhin A.L. Mnogomodal'nye assistivnye sistemy dlya intellektual'nogo zhilogo prostranstva [Multimodal assistive systems for a smart living environment]. *SPIIRAS Proceedings*, 2011, no. 4 (19), pp. 48–64.

Карпов Алексей Анатольевич – доктор технических наук, доцент, ведущий научный сотрудник, Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, 199178, Россия

Железны Милош – PhD, доцент, доцент,Западночешский Университет, Пльзень, 30614, Чехия, zelezny@kky.zcu.cz

Alexey A. Karpov – D.Sc., Associate professor, leading scientific researcher, Saint Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Saint Petersburg, 199178, Russia

Milos Zelezny – PhD, Associate professor, Associate professor, University of West Bohemia, Pilsen, 30614, Czech Republic, zelezny@kky.zcu.cz

Принято к печати 11.07.14
Accepted 11.07.14