УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2016 Том 16 № 3 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2016 Vol. 16 No 3 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.02

АНАЛИЗ МЕТОДОВ МНОГОМОДАЛЬНОГО ОБЪЕДИНЕНИЯ ИНФОРМАЦИИ ДЛЯ АУДИОВИЗУАЛЬНОГО РАСПОЗНАВАНИЯ РЕЧИ

Д.В. Иванько[a], И.С. Кипяткова[b], А.Л. Ронжин[b], А.А. Карпов[b,a]

[a]Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b]Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, 199178, Российская Федерация
Адрес для переписки: karpov@iias.spb.su

Информация о статье
Поступила в редакцию 15.02.16, принята к печати 15.03.16
doi: 10.17586/2226-1494-2016-16-3-387-401
Язык статьи – русский

Ссылка для цитирования: Иванько Д.В., Кипяткова И.С., Ронжин А.Л., Карпов А.А. Анализ методов многомодального объединения информации для аудиовизуального распознавания речи // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 3. С. 387–401. doi: 10.17586/2226-1494-2016-16-3-387-401

Аннотация

В статье представлен аналитической обзор, охватывающий последние результаты, достигнутые в области аудиовизуального объединения (интеграции) многомодальной информации. Рассматриваются основные проблемы и обсуждаются методы их решения. Одной из важнейших задач аудиовизуальной интеграции является понимание того, как именно модальности взаимодействуют и влияют друг на друга. В данной работе этот вопрос рассматривается в контексте аудиовизуальной обработки речи, в особенности распознавания речи. В первой части обзора изложены базовые принципы аудиовизуального распознавания речи, приводится классификация типов аудио- и визуальных признаков речи. Отдельное внимание уделяется систематизации существующих способов и методов объединения аудиовизуальной информации. Во второй части на основе проведенного анализа области исследований приводится сводный список задач и приложений, использующих аудиовизуальное объединение с указанием методов, способов объединения информации и используемых аудио- и видеопризнаков. Предлагается структуризация методов аудиовизуальной интеграции по типам решаемых задач, а также обсуждаются преимущества и недостатки различных подходов. Приведены выводы, предложена оценка будущего развития области. В ходе дальнейших исследований планируется реализация системы аудиовизуального распознавания слитной русской речи с применением современных методов объединения многомодальной информации.

Ключевые слова

аудиовизуальная интеграция, аудиовизуальное распознавание речи, многомодальный анализ, многомодальное объединение, глубокое обучение

Благодарности

Исследование выполнено при финансовой поддержке фонда РФФИ (проект № 15-07-04415-а и 15-07-04322-а) и Совета по грантам Президента РФ (проекты № МД-3035.2015.8 и МК-5209.2015.8).

ANALYSIS OF MULTIMODAL FUSION TECHNIQUES FOR AUDIO-VISUAL SPEECH RECOGNITION

D.V. Ivanko[a], I.S. Kipyatkova[b], A.L. Ronzhin[b], A.A. Karpov[b,a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation
Corresponding author: karpov@iias.spb.su

Article info
Received 15.02.16, accepted 15.03.16
doi: 10.17586/2226-1494-2016-16-3-387-401
Article in Russian

For citation: Ivanko D.V., Kipyatkova I.S., Ronzhin A.L., Karpov A.A. Analysis of multimodal fusion techniques for audio-visual speech recognition. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 3, pp. 387–401. doi: 10.17586/2226-1494-2016-16-3-387-401

Abstract

The paper deals with analytical review, covering the latest achievements in the field of audio-visual (AV) fusion (integration) of multimodal information. We discuss the main challenges and report on approaches to address them. One of the most important tasks of the AV integration is to understand how the modalities interact and influence each other. The paper

addresses this problem in the context of AV speech processing and speech recognition. In the first part of the review we set out the basic principles of AV speech recognition and give the classification of audio and visual features of speech. Special attention is paid to the systematization of the existing techniques and the AV data fusion methods. In the second part we provide a consolidated list of tasks and applications that use the AV fusion based on carried out analysis of research area. We also indicate used methods, techniques, audio and video features. We propose classification of the AV integration, and discuss the advantages and disadvantages of different approaches. We draw conclusions and offer our assessment of the future in the field of AV fusion. In the further research we plan to implement a system of audio-visual Russian continuous speech recognition using advanced methods of multimodal fusion.

Keywords
audio-visual integration, audio-visual speech recognition, multimodal analysis, multimodal fusion, deep learning

Acknowledgements
The research is financially supported by the Russian Foundation for Basic Research (projects No. 15-07-04415-a and 15-07-04322-a) and by the Council for Grants of the President of Russia (projects No. MD-3035.2015.8 and MK-5209.2015.8).

Введение

На сегодняшний день существуют многочисленные области применения, требующие объединения многомодальных данных. Примерами таких областей могут служить биомедицинские приложения (мониторинг интенсивной терапии и медицинских изображений), транспортные системы (умный автомобиль и дорожные системы), мультимедийный анализ (аудиовизуальная идентификация человека, многомодальное взаимодействие с роботом и многомодальный видеопоиск).

Многомодальное объединение – это синергетическое использование информации, полученной из разных модальностей (каналов взаимодействия). Термин «многомодальная интеграция/многомодальное объединение» может относиться к любой стадии процесса интеграции, где присутствует реальная комбинация различных источников информации. Объединение данных имеет смысл, когда данные предоставляют избыточную и дополнительную информацию [1]. Это уменьшает общую неопределенность и способствует повышению точности, с которой признаки воспринимаются системой. Избыточность информации также служит цели повышения надежности системы в случае ошибки или сбоя в исходных сигналах. Дополнительная информация из нескольких модальностей позволяет использовать признаки, которые невозможно однозначно воспринять, имея лишь информацию от каждой модальности в отдельности. Также благодаря параллельной обработке данных несколько модальностей предоставляют более оперативную информацию.

Аудиовизуальный (АВ) анализ является частным случаем многомодального анализа, в котором входными данными являются только аудио- и видеосигналы. Обе модальности взаимосвязаны и содержат дополняющую друг друга информацию. Например, видимость лица улучшает восприятие речи. В работе [2] проводились исследования по изучению взаимоотношения артикуляционных движений, речевой акустики и формы речевого тракта [3], в которых было доказано существование корреляции между этими параметрами. Порождение и восприятие речи человеком является бимодальным. Особенности бимодальной интеграции АВ информации при восприятии речи были проиллюстрированы еще в 1970-е годы посредством «эффекта Мак-Гурка» [4]. Этот эффект можно легко продемонстрировать при помощи видеоряда с одной фонемой (виземой) и звуковой дорожкой с произношением другой фонемы. Часто воспринятая фонема является третьей, средней между этими двумя фонемами. Например, на видео записаны слоги /га-га/, на фонограмме речи – /ба-ба/, а объединенно многие люди воспринимают слоги /да-да/.

Минимальной базовой единицей речи, которая передает лингвистическую информацию, является фонема. Аналогично, основной визуально различимой единицей, используемой в АВ обработке речи и литературе, посвященной человеческому восприятию [5], является визема. Фонемы отражают результат артикуляции, в то время как виземы определяют место артикуляции [6]. Как правило, АВ анализ речи охватывает два основных этапа. На первом этапе из каждой модальности извлекаются соответствующие информативные признаки. Этот шаг полностью зависит от типа используемых модальностей и от области применения. На втором этапе осуществляется интеграция информации, полученной от разных модальностей.

Существует множество приложений, в которых производится объединение аудио и видео, такие как распознавание речи [7–13], распознавание диктора [14–16], биометрическая верификация [17–21], обнаружение события [22–25], слежение за человеком или объектом [26–31], локализация и слежение за активным диктором [32, 33], анализ музыкального контента, распознавание эмоций, видеопоиск, человеко-машинное взаимодействие, обнаружение голосовой активности и разделение источников звукового сигнала [34–36]. Очевидно, что в некоторых приложениях используются изображения лиц, а иногда даже движения всего тела, а не только лица.

В настоящей работе вводятся основные понятия и приводится обзор последних работ по проблематике АВ интеграции речи. Представлены некоторые из задач, возникающих при объединении двух модальностей. Сравниваются различные способы решения таких проблем, предлагаются перспективы будущего развития данного направления исследований. Рассматриваются последние достижения и подходы в этой области.

Базовые принципы аудиовизуального распознавания речи

Важным вопросом при проектировании АВ систем распознавания является то, как правильно интегрировать знания из различных модальностей (в нашем случае аудио и видео), чтобы сохранить необходимую информацию от каждой модальности, но при этом избавиться от недостатков обеих. Общая структура АВ системы распознавания речи приведена на рис. 1.

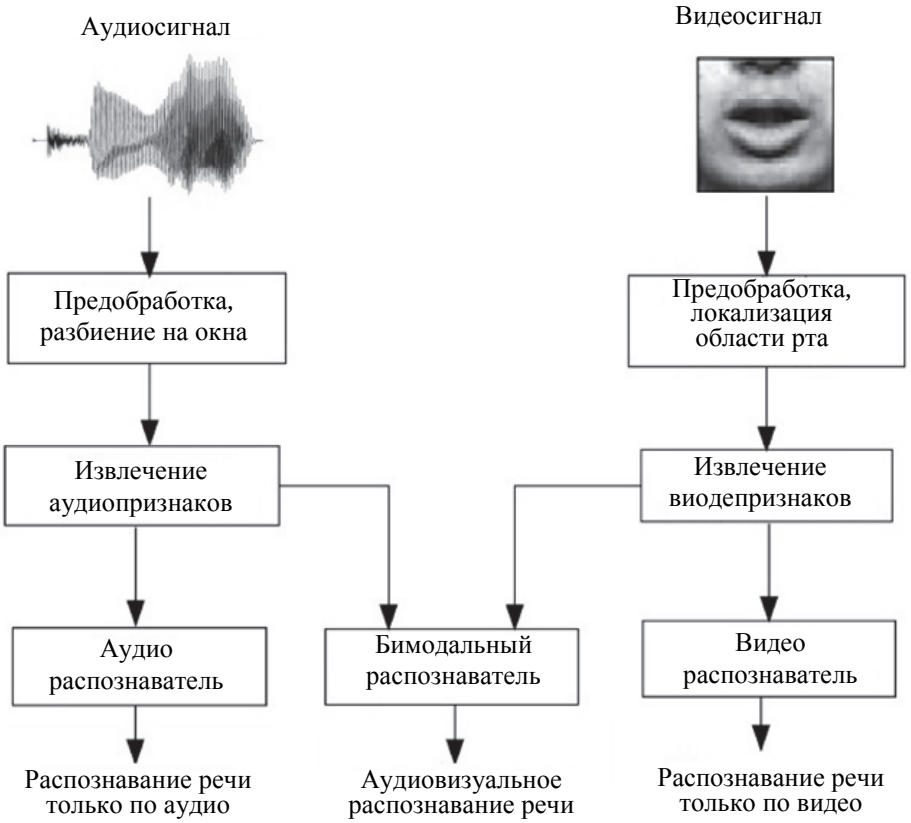


Рис. 1. Общая структура аудиовизуальной системы распознавания речи

При проектировании АВ системы распознавания возникают следующие проблемы.

1. Эффективность каждой модальности не является одинаковой в различных условиях. В некоторых случаях система должна больше полагаться на аудиоинформацию, например, в темном помещении, тогда как в других условиях необходимо больше полагаться на видео, например, в акустически шумном окружении. Другими словами, система должна быть адаптивной к качеству, надежности и достоверности модальностей. Общий подход к достижению этой цели заключается в рассмотрении весов для каждой модальности во время объединения информации. Взвешивание может быть выполнено путем постоянного динамического регулирования весов модальностей в соответствии с качеством тестовых данных [37, 38] или путем расчета некоторых постоянных весов на основе обучающих данных [39]. В тех случаях, когда в обучающих и тестовых данных качество модальностей речи отличается, необходимо использовать динамическое взвешивание. Проблема оценки соответствующих весов для различных условий остается нерешенной, хотя ряд исследователей уделял ей внимание [40].
2. Работа с несколькими модальностями разных типов может стать причиной проблем рассинхронизации информации. Существует два основных типа асинхронности в АВ интеграции. Первый тип возникает благодаря асинхронности аудио- и видеопотоков. Например, визуальные и звуковые признаки речи не обязательно охватывают именно один и тот же промежуток времени. В результате возникает естественная асинхронность между речью и визуальными данными. В АВ распознавании эту проблему относят к сохраняющейся и преждевременной коартикуляции (Preservatory and anticipatory coarticulation) [41]. Другой тип асинхронности связан с разницей между скоростью считывания и временем обработки различных модальностей. Кроме того, объем данных, который требуется для выполнения определенной задачи, зависит от конкретного приложения. Например, этот объем данных будет больше при решении задачи АВ обнаружения событий по сравнению с АВ распознаванием речи. Устранение рассинхронизации модальностей является одной из важнейших проблем в реальных приложениях.
3. На сегодняшний день доступны большие объемы АВ речевых данных, но, в основном, они не размечены и не сегментированы по времени. Процесс разметки данных требует человеческих ресурсов, что

отнимает много времени и средств. Актуально иметь метод интеграции, который способен извлечь пользу из большого количества неразмеченных данных. Использование неразмеченных данных не рассматривается в большинстве обычных AB методов. Однако недавние исследования [7, 8] проводились со сценариями обработки AB данных без учителя и с частичным обучением с учителем. В целом они рассматривали проблему многомодальной обработки как проблему многовидового обучения (Multiview learning). Предлагались новые методы обучения для решения задач отсутствующей разметки, зашумленной модальности и использования частичного обучения. В [42] подробно освещается процесс создания базы данных украинского жестового языка и получения характеристик жестовой речи с целью выделения структуры и связей между мануальными и речевыми компонентами.

Важным шагом перед интеграцией информации является правильное и эффективное представление модальностей (аудио- и видеосигналов) в пространстве признаков. Для аудиосигналов широко используются мел-частотные кепстральные коэффициенты (*Mel Frequency Cepstral Coefficient*, MFCC), коэффициенты линейного предсказания (Linear Prediction Coefficients, LPC), апостериорные признаки фонем, просодические признаки и т.д. С другой стороны, сложной задачей является возможность вычисления соответствующих визуальных признаков из видеосигналов [43]. Для извлечения же визуальных признаков используются подходы, которые могут быть поделены на четыре группы: на основе изображения, движения, геометрии лица и моделей [44].

В качестве визуальных признаков, описывающих параметры губ диктора (визем), могут использоваться две различные системы:

1. пиксельные визуальные признаки, использующие компактное описание графической области губ, например, анализ главных компонент (Principal Component Analysis, PCA) визуальной области губ человека с программным обнаружением области интереса на изображениях с видеокамеры;
2. геометрические визуальные признаки, использующие анализ цветовой дифференциации изображения и описывающие геометрическую форму губ человека: ширина рта, толщина верхней и нижней губ, видимость языка и зубов.

В большинстве случаев после извлечения визуальных признаков применяются методы понижения размерности пространства признаков. Для захвата временной динамики в аудио- и видеопотоках из новых признаков берутся производные первого и второго порядка. В силу того, что темпы аудио- и видеопотоков различаются, требуется этап интерполяции, чтобы представить их с одинаковой скоростью. Хотя в большинстве случаев информация о модальностях объединяется только после извлечения признаков, можно рассмотреть эту информацию во время извлечения признаков.

В работах [45–48] более подробно описаны подходы к извлечению визуальных признаков, используемых в задачах определения контура губ говорящего, структурно-виземного анализа русской речи и др. В публикациях [49, 50] также рассматриваются методы извлечения визуальных признаков в контексте задачи распознавания речи по губам.

Способы объединения аудиовизуальной информации

Объединение модальностей может выполняться на нескольких различных уровнях. Объединение на уровне векторов признаков делается до начала процесса моделирования путем объединения признаков из всех модальностей. Этот способ называется ранней интеграцией. С другой стороны, возможен способ объединения модальностей на уровне принятия решений. В данном случае моделирование каждого канала выполняется раздельно, а затем выходы или решения моделей интегрируются для принятия окончательного решения [51]. Этот способ известен как поздняя интеграция. Кроме того, существует и другой способ, который находится между ранней и поздней интеграцией и называется промежуточной интеграцией (в некоторых источниках его относят к ранней интеграции). Также можно комплексировать два способа интеграции, выполняя объединение одновременно на двух уровнях, что называют гибридным подходом [52]. Далее эти способы описаны более подробно с анализом их преимуществ и недостатков.

Ранняя интеграция. Иллюстрация способа ранней интеграции приведена на рис. 2. Вначале из двух модальностей извлекаются векторы информативных признаков с последующим объединением признаков в один общий вектор. Этот процесс называется интеграцией признаков. Например, объединение (конкатенация) входных векторов признаков в один вектор является одним из простейших способов интеграции признаков. Затем интегральные векторы признаков подаются на вход метода моделирования речи, который формирует решение о гипотезе распознавания. В способе ранней интеграции корреляция между модальностями обнаруживается на уровне признаков, благодаря чему требуется только один процесс моделирования. Это приводит к снижению сложности реализации по сравнению с другими подходами, которые нуждаются в большем количестве процессов моделирования [53]. Однако векторы признаков должны быть преобразованы и масштабированы для того, чтобы сохранить корреляцию в пространстве признаков. Еще одной проблемой является размер интегрального вектора признаков, который может привести к работе в пространстве признаков высокой размерности. Это может затруднить процесс моделирования и уменьшить масштабируемость системы. Для решения этой проблемы могут быть ис-

пользованы такие методы, такие как анализ главных компонент (Principal Components Analysis, PCA) или линейный дискриминантный анализ (Linear Discriminant Analysis, LDA). Кроме того, из-за разной скорости считывания и времени обработки может возникнуть асинхронность между модальностями. Также надо учитывать тот факт, что векторы признаков, объединенные вместе, должны иметь одинаковую длительность по времени (описывать сегменты сигналов одинаковой длины). Стоит отметить, что, несмотря на то, что интеграция признаков является наиболее распространенным способом ранней интеграции, иногда одна модальность может использоваться для определенной инициализации или подготовки системы, а остальная часть задачи выполняется с использованием другой модальности. Например, в [54] для задачи визуального слежения за несколькими дикторами звуковая модальность использовалась только для инициализации системы, благодаря чему ограничивалось пространство поиска визуального детектора лиц, а впоследствии в системе использовалась только визуальная модальность.

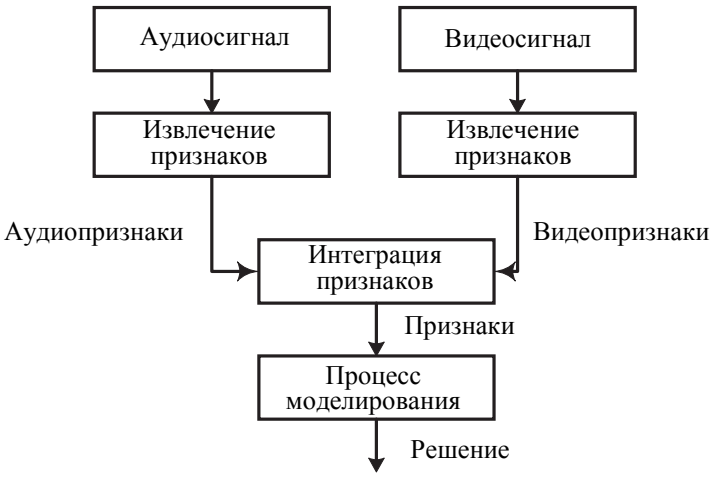


Рис. 2. Способ ранней интеграции аудио- и видеоинформации

Промежуточная интеграция. Способы промежуточной интеграции близки к ранней интеграции. С помощью этих подходов аудио- и видеопризнаки речи предоставляются одному процессу моделирования. Основным отличием этого способа является то, что процесс моделирования специально разработан для обработки нескольких каналов. Он моделирует каждую модальность раздельно с учетом взаимодействия между ними. По сравнению с ранней интеграцией, которая не делает различий между признаками от разных модальностей, промежуточные подходы учитывают разницу между ними, что позволяет им обрабатывать некоторую степень асинхронности между модальностями, а также рассматривать веса для них в различных ситуациях. Основной трудностью промежуточной интеграции является ограничение в выборе методов моделирования, потому что они должны быть разработаны специально для промежуточного интеграционного процесса.

Поздняя интеграция. При таком способе отдельный процесс моделирования принимает признаки одной модальности в качестве входных данных и формирует выходное решение (гипотезу распознавания). Затем решения интегрируются, и финальное решение о гипотезе распознавания принимается блоком интеграции решений. Наиболее простыми методами, используемыми на этом этапе, являются взвешивание, суммирование и голосование. Также могут быть использованы более продвинутые алгоритмы машинного обучения, такие как адаптивное усиление классификаторов (Adaptive Boosting, AdaBoost) и др. В общем виде процесс поздней интеграции представлен на рис. 3.

В позднем способе интеграции выходы процессов моделирования имеют сходные представления гипотез распознавания, и объединить их легче, чем объединить вектора признаков, как это делается при ранней интеграции. Кроме того, обработать асинхронность модальностей легче на уровне принятия решений. Такая система является более масштабируемой по числу модальностей по сравнению со способом ранней интеграции. Еще одно преимущество этого подхода состоит в том, что для каждой конкретной модальности могут быть подобраны соответствующие методы обработки. Например, в задаче AB распознавания речи метод опорных векторов (Support Vector Machines, SVM) используется как процесс моделирования визуальных признаков, в то время как скрытые марковские модели (СММ) используются для речевых сигналов. Основным недостатком способа поздней интеграции является то, что невозможно извлечь непосредственную выгоду из корреляции модальностей на уровне признаков. Кроме того, из-за необходимости раздельного моделирования каждой модальности поздняя интеграция является более сложной в реализации по сравнению с ранней интеграцией.

Как уже говорилось выше, каждый тип интеграции имеет свои плюсы и минусы. Некоторые работы предлагают объединять эти подходы и получить выгоду из преимуществ обоих. Такой подход обычно называют гибридной интеграцией. При этом используется комплексирование методов ранней,

промежуточной и поздней интеграции. Затем для получения окончательного результата распознавания используются решения обеих систем в сочетании с блоком интеграции решений. Таким образом, одновременно можно использовать преимущества как ранней, так и поздней интеграции. Возможные подходы к АВ интеграции речевых модальностей классифицированы на рис. 4.

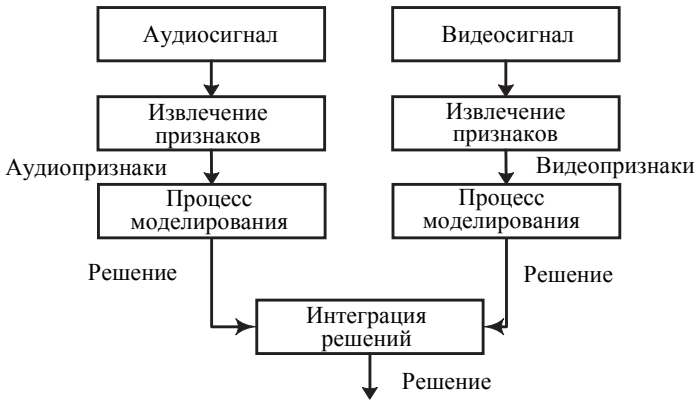


Рис. 3. Способ поздней интеграции аудио- и видеоинформации

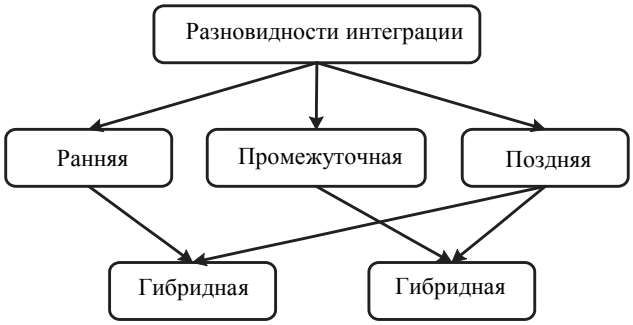


Рис. 4. Классификация подходов к аудиовизуальной интеграции речевых модальностей

Методы аудиовизуального объединения информации

Существует ряд методов, которые были использованы для реализации этапов моделирования и объединения в АВ обработке, таких как метод опорных векторов, графические модели, (например, байесовские сети доверия и СММ), искусственные нейронные сети и алгоритмы оценки (например, фильтрация Калмана). Как правило, эти методы моделирования применимы к различным частям АВ системы распознавания. Рассмотрим некоторые методы более подробно.

Метод опорных векторов (SVM) представляет собой популярный метод моделирования, широко используемый при решении многих задач классификации. В большинстве АВ работ, использовавших SVM, они применялись для независимого моделирования одной модальности. Тем не менее, существуют исследования, особенно с использованием методов поздней интеграции, использующие SVM в качестве метода объединения и интеграции решений, полученных от других компонентов системы. Например, в [34] были проведены исследования по АВ обнаружению определенных событий на видео. При этом некоторые аудио-, визуальные и текстовые данные моделировались раздельно, а затем им давались соответствующие оценки. Полученные оценки затем объединялись для формирования вектора признаков, который подается в SVM с целью обнаружения семантического концепта. Подобная идея используется и в других приложениях, таких как биометрическая идентификация личности, для объединения результатов, полученных от различных компонентов системы (распознавания лиц, верификации диктора, модуля оценки синхронности/корреляции) с последующим использованием SVM в качестве блока интеграции решений.

Еще один популярный метод – динамические байесовские сети (ДБС). Байесовские сети – это вероятностные графические модели, которые представляют собой набор случайных величин с их условными зависимостями. Графическое представление байесовской сети выполняется с помощью ациклических ориентированных графов, в которых вершины представляют собой переменные, а условная зависимость между двумя переменными представлена ребром между соответствующими вершинами [55]. ДБС является байесовской сетью, которая моделирует последовательность наблюдений. Их варианты широко используются в АВ системах, особенно там, где рассматривается временная последовательность, т.е. при обработке речи и видеоанализе. Архитектура ДБС, применяемых для аудиовизуального распознавания речи [11], приведена на рис. 5. Нейронная сеть первоначально предобучается без учителя, после чего настраивается на предсказание вероятностей классов (рис. 5, а). Признаки, извлеченные из слоя «ботлнек»,

получили название DBNF (сокр. от англ. признаки «ботлнек» глубокой нейронной сети, Deep bottleneck features) (рис. 5, б).

Глубокие ДБС были применены в работе [56], где описывается метод обучения пространственно-временных признаков без учителя для аудио- и видеоданных, а также метод обучения их совместного представления. В этом случае сложная структура собирается из составных компонентов различной сложности на разных уровнях абстракции, при этом выход предыдущего слоя предоставляет входные данные текущему слою. Выход каждого слоя объединяет входы посредством пространственных и временных связей. Начиная от первого слоя, входы которого включают в себя сенсорные данные/признаки (например, аудио/акустические признаки, изображение/признаки изображения), составные компоненты каждого слоя становятся все более сложными к выходному слою, который в итоге представляет собой сложные модели образов. В работе [56] на вход первого слоя поступали необработанные данные, включая аудио- и видеоданные. Второй слой являлся скрытым и представлял признаки, извлеченные отдельно из аудио- и видеомодальностей. Третий слой состоял из общих признаков, обученных из аудио- и видеопризнаков. Все признаки во втором и третьем слоях были обучены путем применения метода обучения без учителя. В итоге выходной слой представляет собой некий массив меток, связывающий узлы в слое общих признаков.

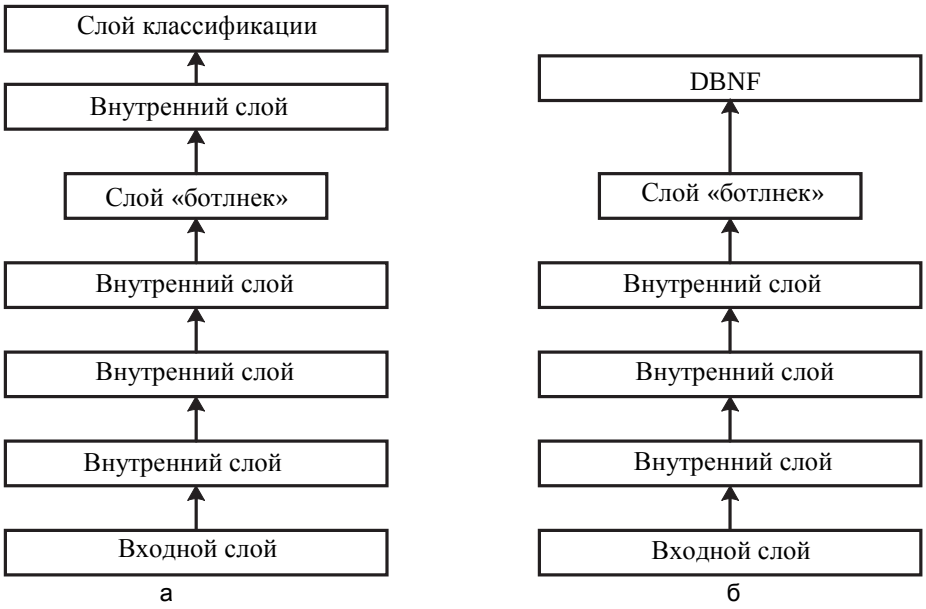


Рис. 5. Архитектуры динамических байесовских сетей: модель классификации (а); извлечение глубоких признаков (б)

В работе [57] предложен подход моделирования на основе двухслойной ДБС, который используется в приложениях видеоанализа для решения следующей задачи: определить, к какому человеку относятся соответствующие аудио- и видеоданные. В первом слое каждая модальность речи независимо моделируется отдельной ДБС. Во втором слое используется новая ДБС для моделирования взаимодействия между этими двумя модальностями. Для оценки параметров ДБС используется алгоритм максимизации ожидания (Expectation maximization, EM). Другие исследователи предлагали использовать многопоточные ДБС для моделирования взаимодействия между модальностями. Например, в [58] для анализа видеоконференций использовалась автоматическая система на основе многопоточных ДБС. Цель – автоматически структурировать данные совещаний, записанных несколькими микрофонами и камерами, в виде последовательности действий встречи, таких как монолог, обсуждение и презентация. В работе было предложено моделировать AV данные совместно с многопоточной ДБС. Общая структура многопоточного ДБС моделирования подробно описана в [59].

В работе [12] использовался подход адаптации объединенных СММ для улучшения распознавания визем по сравнению со стандартной адаптацией и совместным обучением аудиовизуальных СММ. Адаптация объединенных СММ состоит из нескольких независимых этапов, первым из которых является подготовка акустической модели и выравнивание видеокадров. Обучение акустических моделей может быть выполнено с использованием речевых баз данных. Это дает возможность извлечь выгоду из больших аудиокорпусов и создать более точные модели фонем. Как следствие, полученные модели обеспечивают лучшее выравнивание видеокадров. Этот подход не только позволяет получить лучшие акустические модели для конечных аудиовизуальных моделей, но и обеспечивает согласованность, благодаря которой визуальные модели смеси гаусовских распределений (Gaussian mixture model, GMM) добавляются к акустической модели. Данный подход хорошо себя показал при использовании в зашумленных акустических условиях.

Скрытые марковские модели (СММ) могут рассматриваться как простая форма ДБС, которая представляет распределение вероятностей на последовательности наблюдений. Как и ДБС, СММ широко используются в задачах обработки речи и видео.

Многопоточные (multi-stream) СММ (МПСММ) используют два или более отдельных потоков для аудио- и видеонаблюдений с объединением этих наблюдений на каждом сегменте. Архитектура системы аудиовизуального распознавания речи на основе МПСММ [7] приведена на рис. 6.

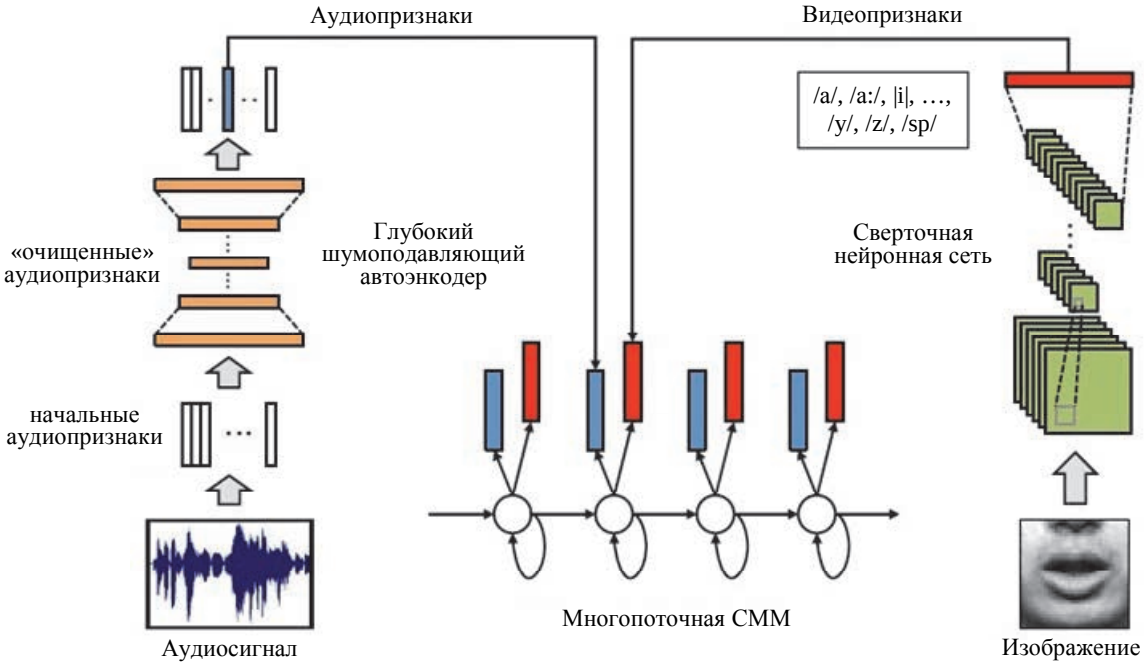


Рис. 6. Общая архитектура системы аудиовизуального распознавания речи на основе многопоточных скрытых марковских моделях

Архитектура системы распознавания основана на двух алгоритмах глубокого обучения: шумоподавляющем автоэнкодере и сверточной нейронной сети для извлечения аудио- и видеопризнаков соответственно. Глубокий шумоподавляющий автоэнкодер обучается предсказывать «очищенные» аудиопризнаки, фильтруя зашумленный входной сигнал. Сверточная нейронная сеть обучается предсказывать произносимые виземы на основе входных изображений области рта. В итоге для интеграции полученных признаков и выдачи результата распознавания используется многопоточная СММ.

Сложность алгоритма декодирования линейно зависит от количества потоков. Этот тип моделирования широко используется в системах АВ распознавания речи. Вместо объединения наблюдений (векторов признаков) на каждом сегменте сигнала в [60] предлагается асинхронная СММ, где две стандартные СММ объединяются на границах модальностей. Акустическая модальность представлена вейвлет-признаками, а визуальная модальность представляется при помощи активных моделей внешнего вида (Active Appearance Model).

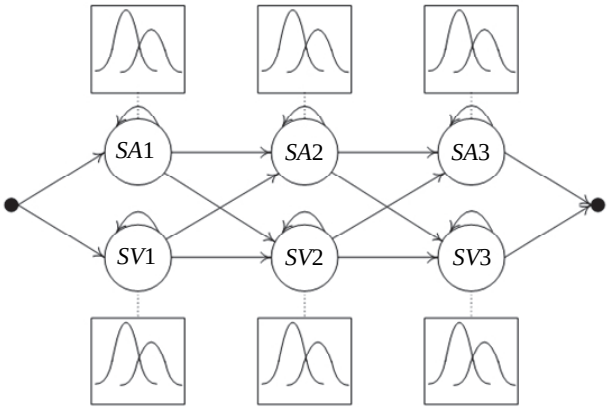


Рис. 7. Топология сдвоенной скрытой марковской модели аудиовизуальной единицы

В другом варианте СММ, называемым сдвоенной (coupled) скрытой марковской моделью (ССММ), множественные потоки информации моделируются с использованием параллельных СММ. На рис. 7 показана топология модели аудиовизуальной единицы речи (пара фонема и визема) с несколькими

состояниями для каждого потока векторов признаков [47]. Кругами обозначены состояния для аудио- и видеомодальностей речи (*SA* и *SV* соответственно), являющиеся скрытыми для наблюдения, а квадратами – смеси нормальных распределений векторов наблюдений в состояниях. Состояния модели в некоторый момент времени t для каждой СММ зависят от скрытых состояний в момент времени $(t-1)$ всех параллельных СММ. Таким образом, общее состояние ССММ определяется совокупностью состояний двух параллельных СММ. Преимущество такой топологии состоит в том, что она позволяет нескольким потокам векторов признаков независимо переходить по состояниям модели, что дает возможность моделировать допустимые временные расхождения в аудио- и видеоданных. В топологии ССММ аудиовизуальных единиц речи применяются по три скрытых состояния на каждый параллельный поток векторов признаков, при этом считается, что первые состояния соответствуют динамическому переходу от предыдущей речевой единицы, третьи – переходу к последующей единице, а вторые состояния объединенной модели (самые длительные) соответствуют стационарному центральному участку элемента речи.

Области применения методов АВ объединения информации

Решаемая задача	**Ссылка**	**Тип аудиопризнаков (речи)**	**Тип видеопризнаков (визуальной информации)**	**Способ объединения**	**Метод объединения информации**
Обнаружение предметов или объектов на видео	[34]	Явно не упоминаются	Множество признаков, включающих цвет, текстуру и ориентированные гистограммы	Гибридная интеграция	SVM
	[35]	MFCC	Низкоуровневые признаки, представляющие цвет, структуру и форму	Поздняя интеграция	SVM, ДБС
	[36]	MFCC	Дискретное косинусное преобразование области лица (DKT) и счет синхронности	Поздняя интеграция	Линейное взвешивание, SVM
	[16]	MFCC	Признаки, основанные на изображении области лица и области губ	Ранняя интеграция	Глубокая нейронная сеть
Биометрическое распознавание личности	[17]	MFCC +Δ+ΔΔ (1 и 2 производные)	Признаки, основанные на форме и внешнем виде	Поздняя интеграция	Взвешенная сумма
	[18]	MFCC	DCT области губ	Поздняя интеграция	SVM
	[19]	MFCC	Признаки на основе формы и интенсивности	Промежуточная интеграция	Асинхронная СММ
	[20]	MFCC +Δ+ΔΔ	Признаки на основе внешнего вида	Поздняя интеграция	Конкатенация, байесовское объединение
	[21]	MFCC +Δ+ΔΔ	Пиксели области лица	Ранняя интеграция	Многовидовое обучение (CCA)
Обнаружение события	[22]	ZCR (Zero-crossing rate), LPC, LFCC (linear frequency cepstral coefficients)	RGB-каналы и определение областей всплеска	Гибридная интеграция	ДБС
	[23]	Уровень волнения говорящего	Множество признаков, таких как двигательная активность и плотность линий поля	Гибридная интеграция	ДБС
	[24]	MFCC, основной тон	Множество признаков на основе слежения за положением головы, губами, бровями, зрачками	Ранняя интеграция	RCCA (Regularized canonical correlation analysis)
	[25]	MFCC	Признаки на основе изменения положения 20 точек на лице	Поздняя интеграция	Интеграция на основе предсказания

Таблица 1. Перечень задач, решаемых с применением методов объединения аудиовизуальной информации

Краткая сводная информация о задачах и прикладных системах, использующих методы АВ объединения, приведена в табл. 1. В каждом из методов АВ признаки использовались совместно с определенным способом интеграции, что также отражено в таблице.

Основная проблема с ССММ и асинхронными СММ состоит в том, что при использовании более двух потоков данных становится проблематичной реализация их алгоритмов обучения. В работе [61] в дополнение к ДБС и СММ для многомодальной интеграции были использованы и другие виды графических моделей, такие как условные случайные поля (Conditional random fields, CRF).

Использование СММ и их вариаций как метода объединения многомодальной информации пользуется большой популярностью при решении задач распознавания речи, что отражено в табл. 2.

Методы объединения нескольких модальностей на основе оценок включают в себя варианты фильтров Калмана и методы фильтрации частиц (Particle filtering) [32]. Фильтрация Калмана – это метод оценки модели пространство–состояние, заключающийся в последовательности наблюдений шумов в течение достаточно долгого времени. Фильтр Калмана представляет собой оптимальный фильтр для одномерных линейных систем с аддитивным гауссовым шумом. Нелинейная версия фильтра Калмана называется расширенным фильтром Калмана и используется для моделирования нелинейных систем. Фильтры частиц используются для моделирования стохастических динамических систем в течение определенного периода времени и также известны как методы последовательностей Монте-Карло. В то время как фильтр Калмана обычно используется для моделирования линейных систем, а расширенный фильтр Калмана используется для нелинейных систем, фильтры частиц являются более подходящими для нелинейных и гауссовых моделей, в частности, для моделей с достаточно большим числом образцов. Эти методы получили популярность при решении задач локализации объектов, объединения данных и в задачах слежения за человеком/объектом (табл. 3). Они могут быть использованы на обоих уровнях объединения – на уровне признаков и на уровне принятия решений.

В литературе можно встретить также дополнительные методы АВ интеграции модальностей, которые зачастую разрабатываются без какой-либо конкретной прикладной задачи. Эти методы интеграции в основном рассматриваются в качестве промежуточных подходов. Например, в [62] предлагается метод, основанный на разреженном представлении для задачи разделения речи разных дикторов, в котором строятся два словаря для выражения избыточного представления аудио- и видеомодальностей. Похожая идея использования двух словарей с целью раздельного моделирования аудио- и видеосигналов с созданием уникального словаря была предложена и в работе [63].

Ссылка	**Тип аудиопризнаков (речи)**	**Тип видеопризнаков (визуальной информации)**	**Способ объединения**	**Метод объединения информации**
[7]	MFCC	Пиксели области губ	Ранняя интеграция	Разреженные RBM (Restricted Boltzmann Machines) и глубокий автоэнкодер
[8]	MFCC и LMFB (Log Mel-Frequency Filterbank)	Пиксели области рта	Промежуточная интеграция	Глубокий автоэнкодер и сверточная нейронная сеть
[9]	MFCC	2D-DCT области губ	Промежуточная интеграция	Сдвоенные СММ
[10]	MFCC	Параметры лицевой анимации (FAP) MPEG-4	Промежуточная интеграция	Многопоточная ДБС
[11]	MFCC	DBVF (Deep Bottleneck Visual Feature)	Промежуточная интеграция	ДБС типа «бутылочное горлышко»
[12]	PLP (Perceptual Linear Prediction)	Изображение области рта + DCT	Гибридная интеграция	Многопоточная СММ
[13]	–	Изображения области рта с нескольких (4) камер	Поздняя интеграция	Многопоточная (4) СММ
[14]	MFCC	Пиксели области рта	Поздняя интеграция	Глубокая нейронная сеть
[15]	MFCC	Пиксели области рта	Гибридная интеграция	Многопоточная СММ
[16]	PLP	60-мерный LDA-вектор признаков	Гибридная интеграция	СММ

Таблица 2. Методы объединения аудиовизуальной информации в задаче распознавания речи

Решаемая задача	Ссылка	Тип аудиопризнаков (речи)	Тип видеопризнаков (визуальной информации)	Способ объединения	Метод объединения информации
Слежение за человеком/ объектом	[26]	MFCC	Пиксели области рта	Промежуточная интеграция	TDNN (Time-Delay Neural Network), ДБС
	[27]	Время задержки прибытия сигнала (TDOA)	Позиция, скорость, размер цели	Поздняя интеграция	Иерархический фильтр Калмана
	[28]	TDOA	Градиент	Ранняя интеграция	Частичная фильтрация
	[29]	TDOA	Координаты	Поздняя интеграция	Частичная фильтрация
	[30]	TOA	Пиксели видеокадров	Промежуточная интеграция	Частичная фильтрация
	[31]	TDOA	Цвет кожи, совпадение контуров, цветовые гистограммы	Поздняя интеграция	Частичная фильтрация
Локализация и слежение за диктором	[32]	TDOA	Позиция диктора	Поздняя интеграция	Расширенный фильтр Калмана
	[33]	MFCC	DCT области рта	Ранняя интеграция	СММ
	[5]	Спектральные компоненты	Мелкомасштабный вид и положение области губ	Промежуточная интеграция	Вероятностная генеративная модель

Таблица 3. Задачи, решаемые с применением методов объединения информации на основе оценок

Заключение

В статье представлен аналитический обзор современных методов объединения информации. Доказывается актуальность их применения не только в задачах распознавания речи, но и в ряде смежных областей, таких как биометрическое распознавание личности, слежение за человеком/объектом, локализация диктора, обнаружение событий на видео и т.д. По результатам анализа выделяется несколько способов интеграции аудиовизуальной информации: ранняя, промежуточная, поздняя и гибридная. Отмечены преимущества и недостатки различных подходов, и наряду с этим даны определения и систематизация современных методов аудиовизуальной интеграции.

Приводится сводный список задач и приложений, использующих аудиовизуальное объединение, с указанием методов, способов объединения информации, используемых аудио- и видеопризнаков. Данный список задач не является исчерпывающим, но дает возможность определить наиболее актуальные области применения и методы интеграции на сегодняшний день. В частности, благодаря этому можно сделать вывод об актуальности использования методов глубокого обучения в данной области. Глубокое обучение, несомненно, улучшит производительность аудиовизуальной интеграции, как это уже было в других областях, которые оно затронуло. Его только начали использовать в области аудиовизуальной интеграции, но первые полученные результаты являются весьма перспективными.

Основываясь на актуальности направления аудиовизуального распознавания речи, в дальнейшем авторами планируется реализация системы аудиовизуального распознавания слитной русской речи с использованием микрофона и высокоскоростной видеокамеры с применением современных методов многомодального объединения информации.

References

1. Katsaggelos A.K., Bahaadini S., Molina R. Audiovisual fusion: challenges and new approaches. *Proc. of the IEEE*, 2015, vol. 103, no. 9, pp. 1635–1653. doi: 10.1109/JPROC.2015.2459017
2. Narayanan S., Alwan A. Noise source models for fricative consonants. *IEEE Transactions on Speech and Audio Processing*, 2000, vol. 8, no. 3, pp. 328–344. doi: 10.1109/89.841215
3. Yehia H., Rubin P., Vatikiotis-Bateson E. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 1998, vol. 26, no. 1–2, pp. 23–43.
4. McGurk H., MacDonald J. Hearing lips and seeing voices. *Nature*, 1976, vol. 264, no. 5588, pp. 746–748.

5. Hershey J., Attias H., Jojic N., Kristjansson T. Audio-visual graphical models for speech processing. *Proc. IEEE International Conference Acoustics, Speech and Signal Processing*, 2004, vol. 5, pp. 649–652.
6. Nock H.J., Iyengar G., Neti C. Speaker localisation using audio-visual synchrony: an empirical study. *Lecture Notes in Computer Science*, 2003, vol. 2728, pp. 488–499.
7. Ngiam J., Khosla A., Kim M., Nam J., Lee H., Ng A.Y. Multimodal deep learning. *Proc. 28th International Conference on Machine Learning*. Bellevue, USA, 2011, pp. 689–696.
8. Noda K., Yamaguchi Y., Nakadai K., Okuno H.G., Ogata T. Audio-visual speech recognition using deep learning. *Application Intelligence*, 2015, vol. 42, no. 4, pp. 722–737. doi: 10.1007/s10489-014-0629-7
9. Nefian A.V., Liang L., Pi X., Liu X., Murphy K. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advanced Signal Processing*, 2002, vol. 2002, no. 11, pp. 1274–1288. doi: 10.1155/S1110865702206083
10. Terry L., Katsaggelos A.K. A phone-viseme dynamic Bayesian network for audio-visual automatic speech recognition. *Proc. 19th International Conference Pattern Recognition*, 2008, art. 4761927.
11. Ninomiya H., Kitaoka N., Tamura S., Iribe Y., Takeda K. Integration of deep bottleneck features for audio-visual speech recognition. *Proc. 16th Annual Conference of the International Speech Communication Association, Interspeech 2015*. Dresden, Germany, 2015, pp. 563–567.
12. Kalantari S., Dean D., Ghaemmaghami H., Sridharan S., Fookes C. Cross database training of audio-visual hidden Markov models for phone recognition. *Proc. 16th Annual Conference of the International Speech Communication Association, Interspeech 2015*. Dresden, Germany, 2015, pp. 553–557.
13. Biswas A., Sahu P.K., Bhowmick A., Chandra M. AAM based features for multiple camera visual speech recognition in car environment. *Procedia Computer Science*, 2015, vol. 57, pp. 614–621. doi: 10.1016/j.procs.2015.07.417
14. Mroueh Y., Marcheret E., Goel V. Deep multimodal learning for audio-visual speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Brisbane, Australia, 2015, pp. 2130–2134. doi: 10.1109/ICASSP.2015.7178347
15. Navarathna R., Dean D., Sridharan S., Lucey P. Multiple cameras for audio-visual speech recognition in an automotive environment. *Computer Speech and Language*, 2013, vol. 27, no. 4, pp. 911–927. doi: 10.1016/j.csl.2012.07.005
16. Marcheret E., Potamianos G., Vopicka J., Goel V. Detecting audio-visual synchrony using deep neural networks. *Proc. 16th Annual Conference of the International Speech Communication Association, Interspeech 2015*. Dresden, Germany, 2015, pp. 548–552.
17. Aleksic P., Katsaggelos A. An audio-visual person identification and verification system using FAPS as visual features. *Proc. ACM Workshop Multimodal User Authentication*, 2003, pp. 80–84.
18. Keating P.A. Underspecification in phonetics. *Phonology*, 1988, vol. 5, no. 2, pp. 275–292.
19. Bengio S. Multimodal authentication using asynchronous HMMs. *Lecture Notes in Computer Science*, 2003, vol. 2688, pp. 770–777.
20. Kanak A., Erzin E., Yemez Y., Tekalp A.M. Joint audio-video processing for biometric speaker identification. *Proc. IEEE International Conference on Acoustic Speech and Signal Processing*. Hong Kong, 2003, vol. 2, pp. 377–380.
21. Chetty G., Wagner M. Audio-visual multimodal fusion for biometric person authentication and liveness verification. *Proc. NICTA-HCSNet Multimodal User Interaction Workshop*, 2006, vol. 57, pp. 17–24.
22. Atrey P.K., Kankanhalli M.S., Jain R. Information assimilation framework for event detection in multimedia surveillance systems. *Multimedia Systems*, 2006, vol. 12, no. 3, pp. 239–253. doi: 10.1007/s00530-006-0063-8
23. Xu H., Chua T.-S. Fusion of AV features and external information sources for event detection in team sports video. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2006, vol. 2, no. 1, pp. 44–67.
24. Shao X., Barker J. Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment. *Speech Communication*, 2008, vol. 50, no. 4, pp. 337–353. doi: 10.1016/j.specom.2007.11.002
25. Petridis S., Rajgarhia V., Pantic M. Comparison of single-model and multiple-model prediction-based audiovisual fusion. *Facial Analysis, Animation and Auditory-Visual Speech Processing, FAAVSP*. Vienna, Austria, 2015, pp. 109–114.
26. Zou X., Bhanu B. Tracking humans using multi-modal fusion. *Proc. IEEE Computer Society Conference Computer Vision and Pattern Recognition Workshops*. San Diego, USA, 2005, pp. 4–11. doi: 10.1109/CVPR.2005.545
27. Talantzis F., Pnevmatikakis A., Polymenakos L.C. Real time audio-visual person tracking. *Proc. IEEE 8th Workshop Multimedia Signal Process*. Victoria, Canada, 2006, pp. 243–247. doi: 10.1109/MMSP.2006.285306

28. Vermaak J., Gangnet M., Blake A., Perez P. Sequential Monte Carlo fusion of sound and vision for speaker tracking. *Proc. IEEE International Conference on Computer Vision*. Vancouver, Canada, 2001, vol. 1, pp. 741–745.
29. Gatica-Perez D., Lathoud G., McCowan I., Odobez J.M., Moore D. Audio-visual speaker tracking with importance particle filters. *Proc. IEEE International Conference on Image Processing*. Barcelona, Spain, 2003, vol. 3, pp. 25–28.
30. Crisan D., Doucet A. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 2002, vol. 50, no. 3, pp. 736–746. doi: 10.1109/78.984773
31. Zotkin D.N., Duraiswami R., Davis L.S. Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, 2002, vol. 2002, no. 11, pp. 1154–1164. doi: 10.1155/S1110865702206058
32. Gehrig T., Nickel K., Ekenel H. K., Klee U., McDonough J. Kalman filters for audio-video source localization. *Proc. IEEE Workshop on Applied Signal Processing to Audio and Acoustics*. New Paltz, USA, 2005, pp. 118–121. doi: 10.1109/ASPAA.2005.1540183
33. Nock H.J., Iyengar G., Neti C. Speaker localisation using audio-visual synchrony: an empirical study. *Lecture Notes in Computer Science*, 2003, vol. 2728, pp. 488–499.
34. Wu Y., Chang K.C., Chang E.Y., Smith J.R. Optimal multimodal fusion for multimedia data analysis. *Proc. 12th ACM International Conference on Multimedia*. New York, 2004, pp. 572–579.
35. Adams W.H., Iyengar G., Lin C.-Y., Naphade M.R., Neti C., Nock H.J., Smith J.R. Semantic indexing of multimedia content using visual, audio, text cues. *EURASIP Journal on Advanced Signal Processing*, vol. 2003, no. 2, pp. 170–185. doi: 10.1155/S1110865703211173
36. Iyengar G., Nock H.J., Neti C. Discriminative model fusion for semantic concept detection and annotation in video. *Proc. 11th ACM International Conference on Multimedia*. Berkeley, USA, 2003, pp. 255–258.
37. Anderson B.D.O., Moore J.B. *Optimal Filtering*. NY, Courier Dover, 2012, 368 p.
38. Estellers V., Gurban M., Thiran J.-P. On dynamic stream weighting for audio-visual speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, vol. 20, no. 4, pp. 1145–1157. doi: 10.1109/TASL.2011.2172427
39. Hsu W.H.-M., Chang S.-F. Generative, discriminative, ensemble learning on multi-modal perceptual fusion toward news video story segmentation. *Proc. IEEE International Conference on Multimedia and Expo*. Taipei, Taiwan, 2004, vol. 2, pp. 1091–1094.
40. Terry L.H., Livescu K., Pierrehumbert J.B., Katsaggelos A.K. Audio-visual anticipatory coarticulation modeling by human and machin. *Proc. 11th Annual Conference of the International Speech Communication Association, Interspeech 2010*. Makuhari, Japan, 2010, pp. 2682–2685.
41. Terry L. *Audio-Visual Asynchrony Modeling and Analysis for Speech Alignment and Recognition. Ph.D. dissertation*. Evanston, USA, Northwestern University, 2011.
42. Kryvonos Iu.G., Krak Iu.V., Barmak O.V., Shkilniuk D.V. Construction and identification of gesture communication elements. *Kibernetika i Sistemny Analiz*, 2013, no. 2, pp. 3–14.
43. Zhou Z., Zhao G., Hong X., Pietikainen M. A review of recent advances in visual speech decoding. *Image and Vision Computing*, 2014, vol. 32, no. 9, pp. 590–605. doi: 10.1016/j.imavis.2014.06.004
44. Dupont S., Luettin J. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2000, vol. 2, no. 3, pp. 141–151. doi: 10.1109/6046.865479
45. Karpov A., Ronzhin A., Kipyatkova I. Designing a multimodal corpus of audio-visual speech using a high-speed camera. *Proc. 11th IEEE Int. Conf. on Signal Processing*. Beijing, China, 2012, pp. 519–522. doi: 10.1109/ICoSP.2012.6491539
46. Karpov A., Kipyatkova I., Zelezny M. A framework for recording audio-visual speech corpora with a microphone and a high-speed camera. *Lecture Notes in Computer Science*, 2014, vol. 8773, pp. 50–57.
47. Karpov A.A. An automatic multimodal speech recognition system with audio and video information. *Automation and Remote Control*, 2014, vol. 75, no. 12, pp. 2190–2200. doi: 10.1134/S000511791412008X
48. Basov O.O., Karpov A.A. Analysis of strategies and methods for multimodal information fusion. *Informatsionno-Upravliaiushchie Sistemy*, 2015, no. 2(75), pp. 7–14.
49. Kovshov E.E., Zavistovskaya T.A. Development of software for testing algorithms design information structures. *Cloud of Science*, 2014, vol. 1, no. 2, pp. 279–291.
50. Krak Yu.V., Ternov A.S. Lipsreading at sign language: synthesis and analysis. *Speech Technology*, 2014, no. 2, pp. 121–131.
51. Snoek C.G., Worring M., Smeulders A.W. Early versus late fusion in semantic video analysis. *Proc. 13th Annual ACM International Conference on Multimedia*. Singapore, 2005, pp. 399–402. doi: 10.1145/1101149.1101236
52. Wu Z., Cai L., Meng H. Multi-level fusion of audio and visual features for speaker identification. *Lecture Notes in Computer Science*, 2005, vol. 3832, pp. 493–499.
53. Atrey P.K., Hossain M.A., Saddik A.E., Kankanhalli M.S. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 2010, vol. 16, no. 6, pp. 345–379. doi: 10.1007/s00530-010-0182-0

54. Barnard M., Koniusz P., Wang W., Kittler J., Naqvi S.M., Chambers J. Robust multi-speaker tracking via dictionary learning and identity modeling. *IEEE Transactions on Multimedia*, 2014, vol. 16, no. 3, pp. 864–880. doi: 10.1109/TMM.2014.2301977
55. *Bayesian Network*. Available at: https://en.wikipedia.org/wiki/Bayesian_network (accessed 20.12.2015).
56. Zhao Y., Wang H., Ji Q. Audio-visual Tibetan speech recognition based on a deep dynamic Bayesian network for natural human robot interaction. *International Journal of Advanced Robotic Systems*, 2012, vol. 9, no. 258, pp. 57–72. doi: 10.5772/54000
57. Noulas A.K., Krose B.J. EM detection of common origin of multi-modal cues. *Proc. 8th ACM International Conference on Multimodal Interfaces*. Banff, Canada, 2006, pp. 201–208. doi: 10.1145/1180995.1181037
58. Dielmann A., Renals S. Automatic meeting segmentation using dynamic Bayesian networks. *IEEE Transactions on Multimedia*, 2007, vol. 9, no. 1, pp. 25–36. doi: 10.1109/TMM.2006.886337
59. Bilmes J.A., Bartels C. Graphical model architectures for speech recognition. *IEEE Signal Processing Magazine*, 2005, vol. 22, no. 5, pp. 89–100. doi: 10.1109/MSP.2005.1511827
60. Bengio S. Multimodal speech processing using asynchronous hidden Markov models. *Information Fusion*, 2004, vol. 5, no. 2, pp. 81–89. doi: 10.1016/j.inffus.2003.04.001
61. Morency L.-P., de Kok I., Gratch J. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 2010, vol. 20, no. 1, pp. 70–84. doi: 10.1007/s10458-009-9092-y
62. Casanovas A.L., Monaci G., Vandergheynst P., Gribonval R. Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 2010, vol. 12, no. 5, pp. 358–371. doi: 10.1109/TMM.2010.2050650
63. Liu Q., Wang W., Jackson P.J.B., Barnard M., Kittler J., Chambers J. Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking. *IEEE Transactions on Signal Processing*, 2013, vol. 61, no. 22, pp. 5520–5535. doi: 10.1109/TSP.2013.2277834

Иванько Денис Викторович – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, denis.ivanko11@gmail.com

Кипяткова Ирина Сергеевна – кандидат технических наук, старший научный сотрудник, Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, 199178, Российская Федерация, kipyatkova@iias.spb.su

Ронжин Александр Леонидович – кандидат технических наук, старший научный сотрудник, Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, 199178, Российская Федерация, ronzhinal@iias.spb.su

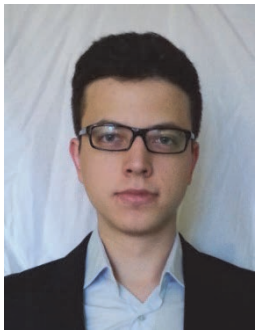
Карпов Алексей Анатольевич – доктор технических наук, доцент, заведующий лабораторией, Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, 199178, Российская Федерация; профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, karpov@iias.spb.su

Denis V. Ivanko – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, denis.ivanko11@gmail.com

Irina S. Kipyatkova – PhD, senior researcher, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation, kipyatkova@iias.spb.su

Alexander L. Ronzhin – PhD, senior researcher, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation, ronzhinal@iias.spb.su

Alexey A. Karpov – D.Sc., Associate professor, Laboratory head, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation; Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, karpov@iias.spb.su



Иванько Денис Викторович – аспирант 1-го года обучения Университета ИТМО по специальности «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей». В 2015 г. окончил магистратуру Университета ИТМО по специальности «Речевые информационные системы». Сотрудник международной лаборатории «Многомодальные биометрические и речевые системы». Автор/соавтор 7 статей, опубликованных в международных и отечественных научных изданиях. Область научных интересов – распознавание речи, обработка аудиовизуальной речи, многомодальные человеко-машинные интерфейсы, многомодальное объединение информации.

Denis V. Ivanko is the 1st year post-graduate student of ITMO University with specialty on mathematical methods and software for computers, systems and networks. He has graduated

from ITMO University in 2015 with specialty on speech information technologies. He is working in the International research laboratory “Multimodal biometric and speech systems”. He has published 7 articles in International and Russian scientific journals and editions. His research interests are: speech recognition, audio-visual speech processing, multimodal human-computer interfaces, multimodal fusion.

Кипяткова Ирина Сергеевна – старший научный сотрудник лаборатории речевых и многомодальных интерфейсов ФГБУН Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН), кандидат технических наук (2011 г.). В 2008 г. окончила Санкт-Петербургский государственный университет аэрокосмического приборостроения (СПбГУАП). С 2007 г. по настоящее время работает в СПИИРАН в лаборатории речевых и многомодальных интерфейсов (до 2008 г. – в группе речевой информатики). Автор/соавтор более 50 статей, опубликованных в международных и отечественных научных изданиях. Область научных интересов – речевые технологии, автоматическое распознавание речи, многомодальные человеко-машинные интерфейсы.

Irina S. Kipyatkova is the senior researcher of the Speech and Multimodal Interfaces Laboratory of St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), PhD (2011). She has graduated from St. Petersburg State University of Aerospace Instrumentation (SUAI) in 2008. She has been working since 2007 in the Speech and Multimodal Interfaces Laboratory (until 2008 in the Speech Informatics Group) of SPIIRAS. She has published more than 50 articles in International and Russian scientific journals and editions. Her research interests are: speech technology, automatic speech recognition, multimodal human-computer interfaces.

Ронжин Александр Леонидович – старший научный сотрудник лаборатории речевых и многомодальных интерфейсов ФГБУН Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН), кандидат технических наук (2013 г.). В 2010 г. окончил Санкт-Петербургский государственный университет аэрокосмического приборостроения (СПбГУАП). С 2008 г. по настоящее время работает в СПИИРАН в лаборатории речевых и многомодальных интерфейсов. Автор/соавтор более 45 статей, опубликованных в международных и отечественных научных изданиях. Область научных интересов – многоканальная обработка аудиовизуальных сигналов, компьютерное зрение, интеллектуальное пространство.

Alexander L. Ronzhin is the senior researcher of the Speech and Multimodal Interfaces Laboratory of St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Ph.D. (2013). He has graduated from St. Petersburg State University of Aerospace Instrumentation (SUAI) in 2010. He has been working since 2008 in the Speech and Multimodal Interfaces Laboratory of SPIIRAS. He has published more than 45 articles in International and Russian scientific journals and editions. His research interests are: multichannel audiovisual signal processing, computer vision, intelligent space.

Карпов Алексей Анатольевич – заведующий лабораторией речевых и многомодальных интерфейсов ФГБУН Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН), профессор кафедры речевых информационных систем Университета ИТМО (по совместительству), доктор технических наук (2014 г.), доцент по специальности (2012 г.). В 2002 г. окончил Санкт-Петербургский государственный университет аэрокосмического приборостроения (СПбГУАП). С 2002 г. по настоящее время работает в СПИИРАН в лаборатории речевых и многомодальных интерфейсов (до 2008 г. – в группе речевой информатики), с 2015 г. возглавляя данную лабораторию, а также с 2014 г. – в Университете ИТМО. Автор/соавтор более 220 статей, опубликованных в международных и отечественных научных изданиях, 3 монографий. Признанный в мире специалист в области речевых технологий и многомодальных интерфейсов. Область научных интересов – речевые технологии, автоматическое распознавание речи, обработка аудиовизуальной речи, многомодальные человеко-машинные интерфейсы.

Alexey A. Karpov is the Head of the Speech and Multimodal Interfaces Laboratory of St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Professor at the Speech Information Systems Department of ITMO University, Dr.Tech.Sc. (2014), Assoc. Professor (2012). He has graduated from St. Petersburg State University of Aerospace Instrumentation (SUAI) in 2002. He has been working since 2002 in the Speech and Multimodal Interfaces Laboratory (until 2008 in the Speech Informatics Group) of SPIIRAS, leading this laboratory since 2015, as well as he is the Professor of ITMO University since 2014. He has published more than 220 articles in International and Russian scientific journals and editions, including 3 monographs. He is a famous International expert in speech technology and multimodal interfaces. His research interests are: speech technology, automatic speech recognition, audio-visual speech processing, multimodal human-computer interfaces.