УДК 004.522
doi:10.31799/1684-8853-2019-2-26-34

Научные статьи
Articles

# Исследование влияния высокоскоростных видеоданных на точность распознавания аудиовизуальной речи

***Д. В. Иванько***[а, б, в], *аспирант, orcid.org/0000-0003-0412-7765, denis.ivanko11@gmail.com*
***Д. А. Рюмин***[а, б], *аспирант, orcid.org/0000-0002-7935-0569*
***А. А. Карпов***[а, б], *доктор техн. наук, доцент, orcid.org/0000-0003-3424-652X*
***М. Железны***[г], *доктор техн. наук, доцент, orcid.org/0000-0003-1695-4370*
[а]*Санкт-Петербургский институт информатики и автоматизации РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*
[б]*Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Кронверкский пр., 49, Санкт-Петербург, 197101, РФ*
[в]*Ульмский университет, Гельмгольцштрассе, 16, 89081, Ульм, Германия*
[г]*Западночешский университет, Университетская ул., 2732/8, 30100, Пльзень, Чехия*

***Введение:*** *эффективность работы современных систем автоматического распознавания речи в тихих акустических условиях достаточно высока и в среднем достигает 90–95 %. Однако в неконтролируемой среде зачастую происходит искажение звукового сигнала, что сильно снижает результирующую точность распознавания. В подобных условиях представляется целесообразным использовать визуальную информацию о речи, так как она не подвержена влиянию акустического шума. На настоящий момент не существует исследований, объективно показывающих зависимость точности распознавания визуальной речи от частоты кадров видео. Также отсутствуют соответствующие аудиовизуальные базы данных для обучения моделей.* ***Цель:*** *сбор представительной базы данных, разработка и исследование автоматической системы аудиовизуального распознавания слитной русской речи.* ***Методы:*** *для распознавания речевых сигналов применяются методы на основе сдвоенных скрытых марковских моделей. Для параметрического представления акустических и визуальных сигналов применяются методы на основе мел-частотных кепстральных коэффициентов и пиксельные признаки, использующие анализ главных компонент.* ***Результаты:*** *исследовались видеоданные с пятью различными скоростями следования кадров: 25, 50, 100, 150 и 200 кадров в секунду. Эксперименты показали положительный эффект от использования высокоскоростной видеокамеры: удалось добиться абсолютного прироста точности на 1,48 % для бимодальной и 3,10 % для одномодальной системы по сравнению со стандартной скоростью записи 25 кадров в секунду. В результате экспериментов с зашумленными данными удалось установить, что бимодальное распознавание речи превосходит по точности распознавания одномодальное, особенно для низких значений ОСШ < 15 дБ. При очень низких значениях ОСШ < 5 дБ акустическая информация становится неинформативной, и наилучшие результаты показывает одномодальная система видеораспознавания речи.* ***Практическая значимость:*** *использование высокоскоростной камеры позволяет улучшить точность и робастность системы распознавания слитной русской речи.*

***Ключевые слова*** *— высокоскоростная видеокамера, аудиовизуальное распознавание речи, шумовые условия, виземы, многомодальное взаимодействие, чтение речи по губам диктора.*

**Для цитирования:** Иванько Д. В., Рюмин Д. А., Карпов А. А., Железны М. Исследование влияния высокоскоростных видеоданных на точность распознавания аудиовизуальной речи. *Информационно-управляющие системы*, 2019, № 2, с. 26–34. doi:10.31799/1684-8853-2019-2-26-34
**For citation:** Ivanko D. V., Ryumin D. A., Karpov A. A., Zelezny M. Measuring the effect of high-speed video data on the audio-visual speech recognition accuracy. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2019, no. 2, pp. 26–34 (In Russian). doi:10.31799/1684-8853-2019-2-26-34

## Введение

Визуальная речь, т. е. видеозапись речевых звуков, играет важную роль в улучшении свойств устойчивости автоматического распознавания речи (Automatic Speech Recognition — ASR).

Эффективность работы современных систем ASR в тихих акустических условиях достаточно высока и в среднем достигает 90–95 %. Однако в неконтролируемой среде зачастую происходит искажение звукового сигнала, что сильно влияет на результирующую точность ASR. В настоящее время использование мультимедийных данных в коммуникационных технологиях получило широкое распространение, и развился новый подход к повышению производительности ASR в акустически неблагоприятных условиях. Он исходит из того, что речь является бимодальной (аудиовизуальной).

Многие эксперименты продемонстрировали, что способность человека понимать речь находится в зависимости от окружающего шума и снижается при его наличии. Для того чтобы повысить разборчивость речи в шумной обстановке, люди начинают читать движения губ собеседника и объединяют эту информацию с речевыми сигналами. После изучения этого феномена идея применения визуальной информации в ASR была изучена в ряде исследований [1, 2]. Видеозапись речевых звуков может быть крайне полезна для

систем ASR в неконтролируемых шумных условиях, так как визуальная модальность не подвержена влиянию акустического шума. Однако, несмотря на серьезные успехи, достигнутые глубоким обучением в области ASR, такие системы по-прежнему не могут приблизиться к уровню компетентности и шумоустойчивости распознавания речи человеком. Существует ряд ограничений, тормозящих прогресс в данной области, в основном связанных с отсутствием аудиовизуальных корпусов и необходимостью сочетания двух областей знаний: автоматического распознавания речи и компьютерного зрения.

Большинство из представленных в литературе экспериментов по аудиовизуальному распознаванию речи было проведено с использованием небольших корпусов, таких как GRID и CUAVE, или неопубликованных корпусов, таких как IBM ViaVoiceTM [3].

В лаборатории речевых и многомодальных интерфейсов СПИИРАН также проводятся исследования по аудиовизуальному распознаванию русской речи. Проведенные эксперименты на собранном мультимедийном корпусе аудиовизуальных данных (RusAVSpeechCorpus, гос. регистрация № 2011620085) показали, что использование стандартной скорости записи видеоданных 25 кадров в секунду зачастую недостаточно для захвата быстрой динамики движений области губ во время слитной речи. То есть некоторые звуки просто смазываются или сливаются на видеоданных, что в конечном итоге приводит к тому, что видеомодальность не только не улучшает точность работы всей системы, но даже ухудшает ее, внося искажения в гипотезы распознавания. Это происходит при достаточно быстром темпе речи говорящего, однако именно этот аспект (распознавание слитной речи) является наиболее важным с практической точки зрения и заслуживает пристального изучения.

Зарубежные исследования по вопросу чтения речи по губам для английского и голландского языков [4] подтверждают эту точку зрения и показывают, что существует некая зависимость между точностью распознавания визуальной речи и частотой следования видеокадров записи. Однако ответов на то, как именно выражается эта зависимость и какую скорость следования видеокадров считать оптимальной для задачи чтения речи по губам, данные исследования не дают.

Повышенная скорость записи видеокадров (высокоскоростные видеокамеры) давно и успешно используются в ряде задач человеко-машинного взаимодействия. К таким задачам можно отнести распознавание микровыражений лиц, распознавание эмоций, определение заболеваний глаз путем отслеживания характера моргания. Тем не менее для задач распознавания речи высокоскоростные камеры по-прежнему считаются слишком ресурсозатратными. Однако уже сейчас многие современные смартфоны оснащены высокоскоростными видеокамерами. К примеру, Apple iPhone X имеет скоростную видеокамеру, позволяющую записывать короткое видео со скоростью 240 кадров в секунду при разрешении $720 \times 480$ пикселей. А новый смартфон Sony Xperia имеет скорость записи до 960 кадров в секунду. Продолжающийся технологический прогресс позволяет с уверенностью предположить, что уже в ближайшем будущем внедрение систем аудиовизуального распознавания речи с высокоскоростными видеокамерами будет являться новой тенденцией в многомодальном распознавании речи. Вышесказанное определяет актуальность и значимость научного исследования, направленного на создание робастной системы аудиовизуального распознавания русской речи с использованием высокоскоростных видеоданных.

## Аудиовизуальная база данных русской речи

Большинство современных систем ASR основаны на вероятностных моделях и методах обработки [5]. Для обучения подобных систем (обучения акустических и визуальных моделей в рамках статистического подхода к распознаванию речи) необходимо иметь представительную речевую базу данных. Сегодня существует целый ряд аудиовизуальных баз данных русской речи, находящихся как в коммерческом, так и в открытом доступе [6, 7]. Тем не менее все они были записаны с использованием стандартной скорости видеозаписи 25 кадров в секунду и поэтому не подходят для нашего исследования. Высокоскоростных баз данных аудиовизуальной русской речи на момент исследования не существовало.

По этим причинам было разработано программное обеспечение и собран корпус аудиовизуальной русской речи с высокоскоростными видеозаписями HAVRUS [8] (High-Speed Recordings of Audio-Visual Russian Speech) (табл. 1), который включает в себя записи 20 дикторов (10 мужчин и 10 женщин). Средний возраст дикторов составляет 22 года, все они являются нормативными носителями русского языка. Каждый из дикторов произнес 200 подобранных фраз на русском языке: 130 фраз для обучения были взяты из фонетически представительных текстов и являлись общими для всех говорящих, 70 фраз для тестирования отличались для каждого диктора и являлись номерами телефонов.

Продолжительность речевых данных каждого диктора составляет приблизительно 15–20 мин

■ ***Таблица 1***. Общие характеристики речевой базы данных HAVRUS

■ ***Table 1***. General characteristics of the HAVRUS database

| Параметр | Значение |
|---|---|
| Количество дикторов | 20 |
| Количество фраз у каждого диктора | 200 |
| Аудиоданных на каждого диктора, мин | 15–20 |
| Общая длительность аудиоданных, ч | ≈6 |
| Процент чистой речи в записях, % | >80 |
| Частота дискретизации, кГц | 44,1 |
| Квантование сигнала, бит | 16 |
| Отношение сигнал/шум (ОСШ), дБ | >35 |
| Тип микрофона | Oktava MK-012 |
| Тип камеры | JAI Pulnix RMC-6740 |
| Тип объектива | KOWA LM6NCM |
| Разрешение изображения, пикселей | 640 × 480 |
| Общий объем данных, ТБ | ≈5,5 |

слитной речи. Видеоданные имеют частоту кадров 200 кадров в секунду, несжатые данные в RAW-формате.

## Обработка аудиовизуальной русской речи

### Архитектура системы аудиовизуального распознавания русской речи

Архитектура разработанной системы аудиовизуального распознавания русской речи на базе микрофона и высокоскоростной видеокамеры показана на рис. 1. Выходными данными является результат распознавания в текстовой форме, который выводится в окне ПО AVSpeechRecognition. В режиме офлайн-распознавания также в окно программы будет выведено время работы программы, длительность видео, скорость обработки видеофайлов и точность распознавания.

### Методы аудиовизуальной обработки речи

На этапе моделирования и объединения аудиовизуальных данных используется несколько распространенных методов, таких как метод опорных векторов, байесовские сети доверия, скрытые марковские модели (СММ), нейронные сети, оценочные алгоритмы и пр. [9–11].

В настоящей работе применяется хорошо себя зарекомендовавшая технология сдвоенных скрытых марковских моделей (ССММ) по причине того, что они способны учитывать естественную асинхронность движений губ и речевой информации, обеспечивая синхронизацию аудиовизуальной речи на границах слов.

*Объединение аудиовизуальных модальностей речи.*

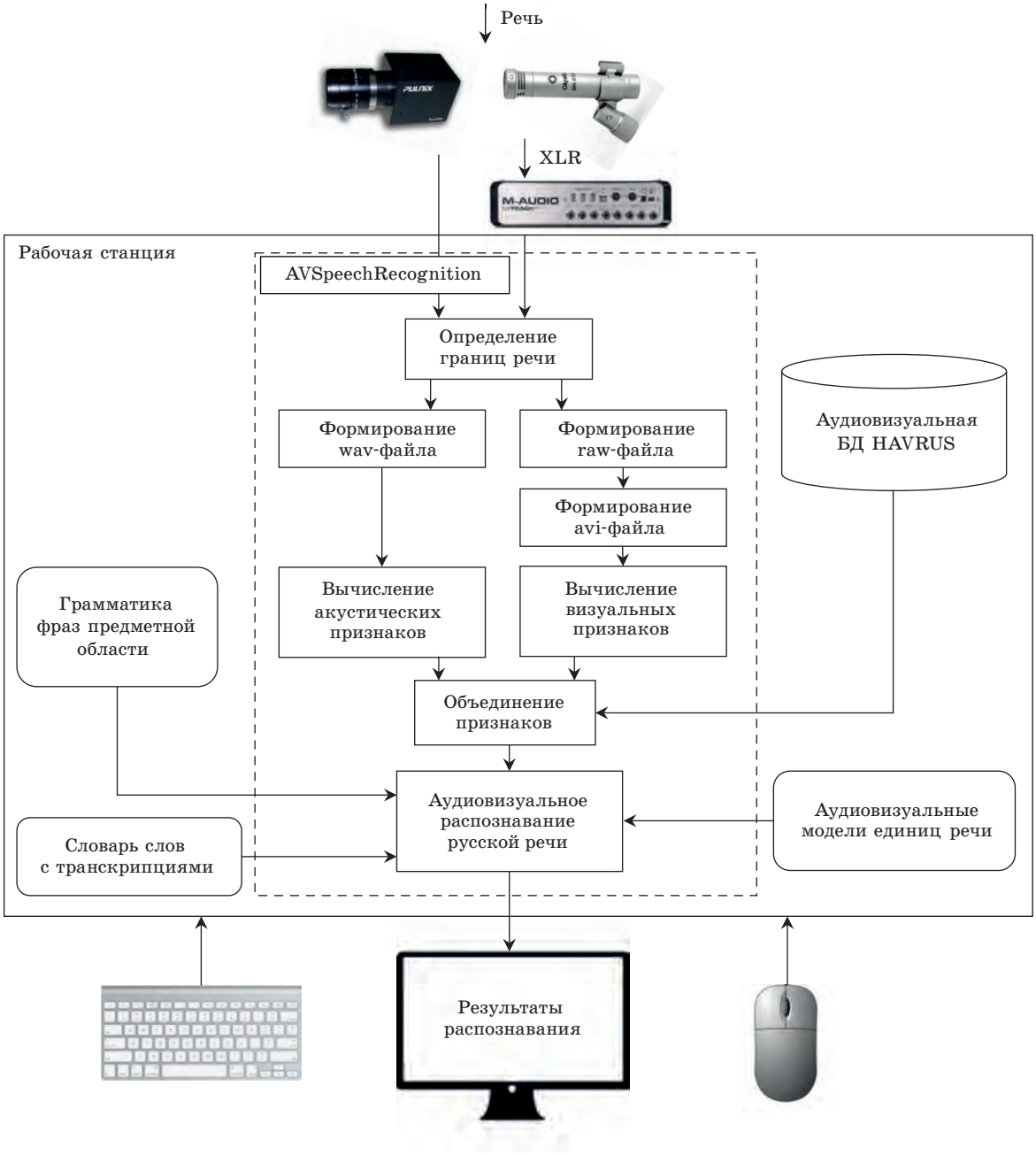
Скрытую марковскую модель можно рассматривать как простую форму байесовской сети доверия, которая представляет распределения вероятностей по последовательностям наблюдений. Как и байесовские сети доверия, СММ широко эксплуатируются в задачах обработки речи и видео. Одной из форм СММ являются многопоточные модели, которые имеют в своей структуре два отдельных потока для аудио- и видеонаблюдений, объединяя эти наблюдения на каждом кадре. Сложность декодирования такого алгоритма является линейной и зависит от количества потоков. В ССММ параллельные потоки моделируются с использованием параллельных СММ, где каждое состояние СММ может проходить в асинхронной области, но должно оставаться привязанным к границам модели [12]. Основная проблема ССММ заключается в том, что их алгоритмы обучения становятся неразрешимыми при наличии более двух потоков (модальностей) [13].

В работе [14] был предложен метод адаптации сдвоенных СММ, применение которого позволяет увеличить точность распознавания визем по сравнению со стандартным методом совместного обучения аудиовизуальных СММ. Его суть заключается в раздельном обучении акустических и визуальных моделей, причем акустическая модель может быть обучена в том числе и с привлечением сторонних баз данных. Объем аудиоданных в современных аудиовизуальных базах не является достаточно большим, и привлечение дополнительных аудиоданных позволяет создать лучшие модели фонем. В итоге это приводит к лучшему выравниванию видеокадров. Этот подход обеспечивает не только лучшие акустические модели для конечных аудиовизуальных моделей, но и лучшую согласованность, благодаря которой лучшие визуальные модели будут добавлены к акустической модели. Данный подход хорошо себя показал при использовании в зашумленных условиях.

### Параметрическое представление аудиовизуальной речи

Для параметрического описания аудиосигнала рассчитываются мел-частотные кепстральные коэффициенты (Mel-Frequency Cepstral Coefficients) с их первой и второй производными.

В отличие от аудио, не существует стандартизированного набора визуальных признаков.




■ ***Рис. 1.*** Архитектура автоматической системы многомодального распознавания русской речи
■ ***Fig. 1.*** Architecture of the automatic system of multimodal Russian speech recognition

Большинство исследователей применяют различные алгоритмы для их извлечения [15–17]. В качестве визуальных признаков, описывающих форму губ человека (визем), были использованы пиксельные визуальные признаки на основе метода анализа главных компонент (Principal Component Analysis) визуальной области губ человека. Визуальные признаки вычисляются в несколько этапов. Во-первых, на видеокадрах происходит поиск области лица человека каскадным классификатором по методу Adaptive Boosting (адаптивное усиление классификаторов), который основан на алгоритме Виолы — Джонса. Результатом подобной обработки является обнаружение области интереса (области рта диктора). После чего выполняется цифровая обработка изображения, включающая в себя следующие этапы: нормализацию области губ до изображения размером 32 × 32 пикселей и отображение в 32-мерный вектор признаков по методу анализа главных компонент [18].

**Классы визем русской речи**

Согласно нашим предыдущим исследованиям [19], наибольшая точность распознавания русской речи достигалась при использовании 20 визуально различимых единиц речи (визем) (табл. 2).

Этот параметр зависит от языка [20, 21], и для русского использовалось 10–14 визем в различных работах. Однако при внедрении высокоскоростной видеокамеры удается намного лучше отследить быструю динамику движения губ в слитной речи.

## Результаты экспериментов

В целях проведения экспериментов было реализовано несколько автоматических систем распознавания речи, как многомодальных (аудиовизуальных), так и одномодальных. Все системы являлись дикторозависимыми с малым словарем распознавания. Затем была проведена их экспериментальная проверка с применением собранного аудиовизуального корпуса русской речи с высокоскоростными видеозаписями HAVRUS. При этом ОСШ изменялось в пределах от 0 до 40 дБ. Модели сравнивались по количественному показателю точности распознавания слов слитной речи WRR.

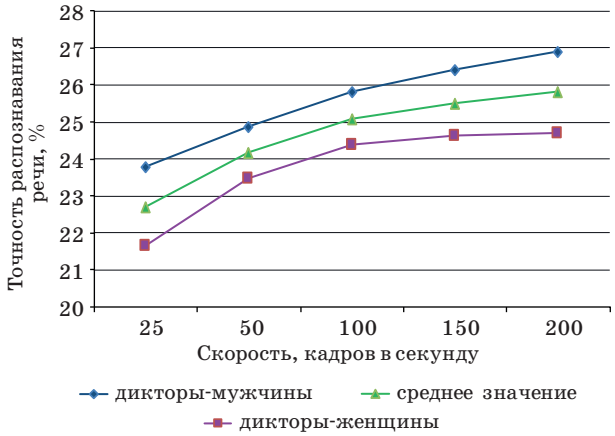
**Распознавание визуальной речи**

В настоящих экспериментах исследовались видеоданные с пятью различными скоростями следования кадров: 25, 50, 100, 150 и 200 кадров в секунду.

■ ***Таблица 2***. Классы визем и их соответствие фонемам русской речи

■ ***Table 2***. Viseme classes and their correspondence to the phonemes of the Russian speech

| Класс виземы | Соответствующие фонемы русской речи | Класс виземы | Соответствующие фонемы русской речи |
|---|---|---|---|
| V1 | Тишина (пауза) | V11 | э! |
| V2 | а, а! | V12 | ы, ы! |
| V3 | и, и! | V13 | у, у! |
| V4 | о! | V14 | э |
| V5 | б, б’, п, п’ | V15 | с, с’, з, з’, ц |
| V6 | ф, ф’, в, в’ | V16 | й |
| V7 | ш, щ | V17 | х, х’ |
| V8 | л, л’, р, р’ | V18 | ч |
| V9 | д, д’, т, т’, н, н’ | V19 | м, м’ |
| V10 | г, г’, к, к’ | V20 | ж |




■ ***Рис. 2***. Изменение пословной точности распознавания речи при увеличении скорости следования видеокадров в системе чтения русской речи по губам

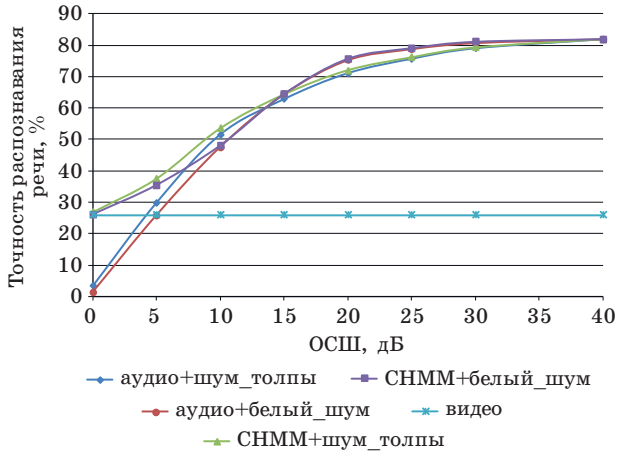
■ ***Fig. 2***. Word recognition rate on increasing speed of video frames rate in the Russian lip reading system

Пословная точность распознавания речи (Word Recognition Rate — WRR) для всех 20 дикторов показана в зависимости от частоты кадров (рис. 2). Как видно из рисунка, WRR значительно увеличивается (около 1,5 %) при увеличении частоты следования видеокадров от 25 до 50. При дальнейшем увеличении частоты до 100 кадров в секунду WRR продолжает расти, однако темп роста замедляется (всего 0,91 % за дополнительные 50 кадров в секунду). Эта тенденция продолжает сохраняться и при увеличении до 150 кадров в секунду — WRR увеличивается еще на 0,42 %, но при увеличении до 200 кадров в секунду прирост точности составил всего 0,3 %.

**Аудиовизуальное распознавание речи**

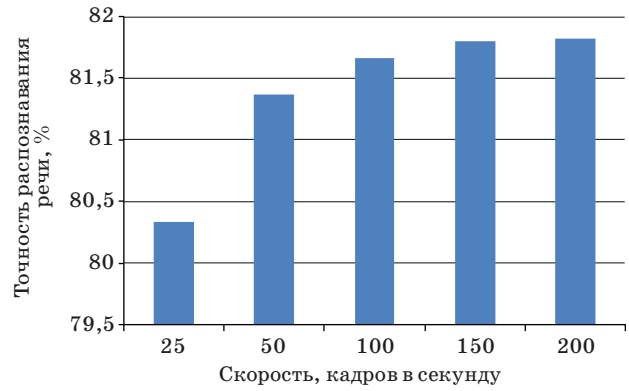
В ходе экспериментов тестовые данные для всех дикторов зашумлялись двумя видами шумов: широкополосным белым шумом и «шумом толпы» — одновременным говорением нескольких людей [22]. Интенсивность аддитивного шума изменялась в пределах от 0 до 40 дБ с шагом 5 дБ.

Результаты экспериментов по проверке трех систем распознавания (одномодальных систем с распознаванием только аудиосигнала или видеосигнала, бимодальной системы ССММ) представлены на рис. 3. Анализ результатов показывает, что бимодальное распознавание речи превосходит по точности распознавания слов одномодальное распознавание, что особенно очевидно для низких значений ОСШ < 15 дБ. При очень низких значениях ОСШ < 5 дБ акустическая информация становится неинформативной, и наилучшие результаты показывает одномодальная система видеораспознавания речи.




■ ***Рис. 3.*** Зависимость пословной точности распознавания речи от ОСШ аудиосигнала для различных конфигураций системы аудиовизуального распознавания речи

■ ***Fig. 3.*** The dependence of word recognition rate on SNR for various configurations of the audio-visual speech recognition system




■ ***Рис. 4.*** Пословная точность аудиовизуального распознавания русской речи при различных значениях скорости следования видеоданных

■ ***Fig. 4.*** WRR values for audio-visual Russian speech recognition with various fps of video data

Наилучшие результаты распознавания достигаются объединенной бимодальной системой. Так, в сильно зашумленных акустических условиях вес аудиомодальности минимизируется, и система полагается только на результат видеомодальности (при ОСШ < 10 дБ). Однако при ОСШ > 10 дБ видеомодальность уже не может обеспечить значительного прироста точности распознавания, поэтому ее вес уменьшается, а аудиомодальности — увеличивается.

При правильном использовании этого метода аудиовизуальная система становится более робастной к акустическим шумам, и появляется возможность добиться наилучших результатов распознавания речи при любых акустических условиях путем изменения весов модальностей. Для повышения качества распознавания речи в реальных условиях применения этот метод следует сочетать с различными методами шумоочистки.

Другая серия проведенных экспериментов была ориентирована на исследование влияния скорости следования видеокадров на точность распознавания бимодальной системы. На рис. 4 представлены результаты работы такой системы при различных уровнях скорости следования кадров в секунду.

Пословная точность распознавания речи достигает своего максимального значения 81,82 % при 200 кадрах в секунду. Это на 1,48 % выше, чем при использовании стандартной скорости записи видеоданных 25 кадров в секунду.

## Заключение

В работе представлены результаты экспериментов по распознаванию русской речи и чтению речи по губам, полученные при помощи аудиовизуальной системы распознавания речи на основе сдвоенных скрытых марковских моделей. Эксперименты показали положительный эффект от использования высокоскоростной видеокамеры: при использовании высокоскоростной камеры удалось добиться абсолютного прироста точности на 1,48 % для бимодальной и 3,10 % для одномодальной системы по сравнению со стандартной скоростью записи 25 кадров в секунду. Эксперименты с системой видеораспознавания речи показали улучшение WRR до 7,28 % для некоторых дикторов.

Основываясь на проведенных исследованиях, можно сделать вывод, что внедрение высокоскоростной камеры позволяет улучшить точность и робастность системы распознавания слитной русской речи. Согласно нашим наблюдениям, белый шум уменьшает точность распознавания системы сильнее, чем «шум толпы». Однако белый шум достаточно легко поддается шумоочистке, тогда как задача очистки «шума толпы» по-прежнему остается большой и до конца нерешенной проблемой. В силу того, что «шум толпы» довольно часто встречается в реальных условиях применения, очень важно иметь надежную систему распознавания речи для таких условий. По нашему мнению, это может быть достигнуто с использованием аудиовизуальной системы распознавания речи на основе высокоскоростной видеокамеры.

Данное исследование проводится при поддержке фонда РФФИ (проекты № 18-37-00306, 16-37-60100, 18-07-01407), Правительства РФ (грант № 08-08), а также бюджетной темы № 0073-2019-0005.

## Литература


1. **Katsaggelos K., Bahaadini S., Molina R.** Audiovisual fusion: challenges and new approaches. *Proc. of the IEEE*, 2015, vol. 103, no. 9, pp. 1635–1653.
2. **Zhou Z., Zhao G., Hong X., Pietikainen M.** A review of recent advances in visual speech decoding. *Proc. of the Image and Vision Computing*, 2014, vol. 32, pp. 590–605.
3. **Ivanko D., Karpov A., Ryumin D., Kipyatkova I., Saveliev A., Budkov V., Zelezny M.** Using a high-speed video camera for robust audio-visual speech recognition in acoustically noisy conditions. *Intern. Conf. on Speech and Computer (SPECOM)*, 2017, pp. 757–766.
4. **Chitu A. G., Driel K., Rothkrantz L. J. M.** Automatic lip reading in the Dutch language using active appearance models on high speed recordings. *Text, Speech and Dialogue, Springer LNCS (LNAI)*, 2010, vol. 6231, pp. 259–266.
5. **Rajavel R., Sathidevi P. S.** Adaptive reliability measure and optimum integration weight for decision fusion audio-visual speech recognition. *Journal of Signal Processing Systems*, 2012, vol. 68, no. 1, pp. 83–93.
6. **Stewart D., Seymour R., Pass A., Ming J.** Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Transactions on Cybernetics*, Feb. 2014, vol. 44, no. 2, pp. 175–184.
7. **Abhishek N., Prasanta K. G.** PRAV: a phonetically rich audio visual corpus. *Proc. of the Interspeech*, 2017, pp. 3747–3751.
8. **Verkhodanova V., Ronzhin A., Kipyatkova I., Ivanko D., Karpov A., Železný M.** HAVRUS corpus: high-speed recordings of audio-visual russian. *Intern. Conf. on Speech and Computer (SPECOM)*, 2016, vol. 9811, pp. 338–345.
9. **Shivappa S. T., Trivedi M. M., Rao B. D.** Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey. *Proc. of IEEE*, 2010, vol. 98, no. 10, pp. 1692–1715.
10. **Abdelaziz A. H., Kolossa D.** Dynamic stream weight estimation in coupled HMM-based audio-visual speech recognition using multilayer perceptrons. *Proc. of the Interspeech*, 2014, pp. 1144–1148.
11. **Huang J., Kingsbury B.** Audio-visual deep learning for noise robust speech recognition. *Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7596–7599.
12. **Graves A., Mohamed A., Hinton G.** Speech recognition with deep recurrent neural networks. *Proc. of the IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.
13. **Shiell D. J., Terry L. H., Aleksic P. S., Katsaggelos A. K.** *Audio-visual and visual-only speech and speaker recognition: Issues about theory, system design and implementation*. In: *Visual Speech Recognition: Lip Segmentation and Mapping*. IGI Global, 2009. Pp. 1–38.
14. **Abdelaziz A. H., Zeiler S., Kolossa D.** A new EM estimation of dynamic stream weights for coupled-HMM-based audio-visual ASR. *Proc. of IEEE Intern. Conf. on Acoustic Speech and Signal Processing (ICASSP)*, 2014, pp. 54–62.
15. **Kumar S., Bhuyan M. K., Chakraborty B. K.** Extraction of texture and geometrical features from informative facial regions for sign language recognition. *Journal of Multimodal User Interfaces (JMUI)*, 2017, vol. 11, no. 2, pp. 227–239.
16. **Lan Y., Theobald B., Harvey E., Ong E., Bowden R.** Improving visual features for lip-reading. *Proc. of Auditory-Visual Speech Processing (AVSP)*, 2010, pp. 142–147.
17. **Хафизов Р. Г., Яранцева Т. В.** Оценка геометрических искажений контуров изображений губ в системах визуального ввода информации. *Информационно-управляющие системы*, 2017, № 4, с. 2–6. doi:10.15217/issn1684-8853.2017.4.2
18. **Кухарев Г. А., Каменская Е. И., Матвеев Ю. Н., Щеголева Н. Л.** *Методы обработки и распознавания изображений лиц в задачах биометрии*. СПб., Политехника, 2013. 388 с.
19. **Ivanko D., Karpov A., Fedotov D., Kipyatkova I., Ryumin D., Ivanko Dm., Minker W., Zelezny M.** Multimodal speech recognition: increasing accuracy using high speed video data. *Journal of Multimodal User Interfaces*, 2018, vol. 12, no. 4, pp. 319–328.
20. **Иванько Д., 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(103), с. 387–401.
21. **Estellers V., Gurban M., Thiran J.** 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.
22. **Stewart D., Seymour R., Pass A., Ming J.** Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Transactions on Cybernetics*, 2013, vol. 44, no. 2, pp. 175–184.

UDC 004.522
doi:10.31799/1684-8853-2019-2-26-34

**Measuring the effect of high-speed video data on the audio-visual speech recognition accuracy**

D. V. Ivanko[a,b,c], Post-Graduate Student, orcid.org/0000-0003-0412-7765, denis.ivanko11@gmail.com
D. A. Ryumin[a,b], Post-Graduate Student, orcid.org/0000-0002-7935-0569
A. A. Karpov[a,b], Dr. Sc., Tech., Associate Professor, orcid.org/0000-0003-3424-652X
M. Zelezny[d], PhD, Associate Professor, orcid.org/0000-0003-1695-4370
[a]Saint-Petersburg Institute for Informatics and Automation of the RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics, 49, Kronverkskii Pr., 197101, Saint-Petersburg, Russian Federation
[c]Universität Ulm, 16, Helmholtzstraße, 89081, Ulm, Germany
[d]University of West Bohemia (UWB), 2732/8, Univerzitní ul., 301 00, Plzeň, Czech Republic

**Introduction:** The effectiveness of modern automatic speech recognition systems in quiet acoustic conditions is quite high and reaches 90–95%. However, in noisy uncontrolled environment, acoustic signals are often distorted, which greatly reduces the resulting recognition accuracy. In adverse conditions, it seems appropriate to use the visual information about the speech, as it is not affected by the acoustic noise. At the moment, there are no studies which objectively reflect the dependence of visual speech recognition accuracy on the video frame rate, and there are no relevant audio-visual databases for model training. **Purpose:** Improving the reliability and accuracy of the automatic audio-visual Russian speech recognition system; collecting representative audio-visual database and developing an experimental setup. **Methods:** For audio-visual speech recognition, we used coupled hidden Markov model architectures. For parametric representation of audio and visual features, we used mel-frequency cepstral coefficients and principal component analysis-based pixel features. **Results:** In the experiments, we studied 5 different rates of video data: 25, 50, 100, 150, and 200 fps. Experiments have shown a positive effect from the use of a high-speed video camera: we achieved an absolute increase in accuracy of 1.48% for a bimodal system and 3.10% for a unimodal one, as compared to the standard recording speed of 25 fps. During the experiments, test data for all speakers were added with two types of noise: wide-band white noise and "babble noise". Analysis shows that bimodal speech recognition exceeds unimodal in accuracy, especially for low SNR values <15 dB. At very low SNR values <5 dB, the acoustic information becomes non-informative, and the best results are achieved by a unimodal visual speech recognition system. **Practical relevance:** The use of a high-speed camera can improve the accuracy and robustness of a continuous audio-visual Russian speech recognition system.

**Keywords** — high-speed video camera, audio-visual speech recognition, noisy conditions, visemes, multimodal processing, lip-reading.

**For citation:** Ivanko D. V., Ryumin D. A., Karpov A. A., Zelezny M. Measuring the effect of high-speed video data on the audio-visual speech recognition accuracy. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2019, no. 2, pp. 26–34 (In Russian). doi:10.31799/1684-8853-2019-2-26-34

**References**

1. Katsaggelos K., Bahaadini S., Molina R. Audiovisual Fusion: Challenges and New Approaches. *Proc. of the IEEE*, 2015, vol. 103, no. 9, pp. 1635–1653.
2. Zhou Z., Zhao G., Hong X., Pietikainen M. A review of recent advances in visual speech decoding. *Proc. of the Image and Vision Computing*, 2014, vol. 32, pp. 590–605.
3. Ivanko D., Karpov A., Ryumin D., Kipyatkova I., Saveliev A., Budkov V., Zelezny M. Using a high-speed video camera for robust audio-visual speech recognition in acoustically noisy conditions. *Intern. Conf. on Speech and Computer (SPECOM)*, 2017, pp. 757–766.
4. Chitu A. G., Driel K., Rothkrantz L. J. M. Automatic lip reading in the Dutch language using active appearance models on high speed recordings. *Text, Speech and Dialogue, Springer LNCS (LNAI)*, 2010, vol. 6231, pp. 259–266.
5. Rajavel R., Sathidevi P. S. Adaptive reliability measure and optimum integration weight for decision fusion audio-visual speech recognition. *Journal of Signal Processing Systems*, 2012, vol. 68, no. 1, pp. 83–93.
6. Stewart D., Seymour R., Pass A., Ming J. Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Transactions on Cybernetics*, Feb. 2014, vol. 44, no. 2, pp. 175–184.
7. Abhishek N., Prasanta K. G. PRAV: a phonetically rich audio visual corpus. *Proc. of the Interspeech*, 2017, pp. 3747–3751.
8. Verkhodanova V., Ronzhin A., Kipyatkova I., Ivanko D., Karpov A., Železný M. HAVRUS Corpus: High-Speed Recordings of Audio-Visual Russian Speech *Intern. Conf. on Speech and Computer (SPECOM)*, 2016, vol. 9811, pp. 338–345.
9. Shivappa S. T., Trivedi M. M., Rao B. D. Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey. *Proc. of IEEE*, 2010, vol. 98, no. 10, pp. 1692–1715.
10. Abdelaziz A. H., Kolossa D. Dynamic stream weight estimation in coupled HMM-based audio–visual speech recognition using multilayer perceptrons. *Proc. of the Interspeech*, 2014, pp. 1144–1148.
11. Huang J., Kingsbury B. Audio-visual deep learning for noise robust speech recognition. *Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7596–7599.
12. Graves A., Mohamed A., Hinton G. Speech recognition with deep recurrent neural networks. *Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.
13. Shiell D. J., Terry L. H., Aleksic P. S., Katsaggelos A. K. *Audio-visual and visual-only speech and speaker recognition: Issues about theory, system design and implementation*. In: *Visual Speech Recognition: Lip Segmentation and Mapping*. IGI Global, 2009. Pp. 1–38.
14. Abdelaziz A. H., Zeiler S., Kolossa D. A new EM estimation of dynamic stream weights for coupled-HMM-based audio-visual ASR. *Proc. of IEEE Intern. Conf. on Acoustic Speech and Signal Processing (ICASSP)*, 2014, pp. 54–62.
15. Kumar S., Bhuyan M. K., Chakraborty B. K. Extraction of texture and geometrical features from informative facial regions for sign language recognition. *Journal of Multimodal User Interfaces (JMUI)*, 2017, vol. 11, no. 2, pp. 227–239.
16. Lan Y., Theobald B., Harvey E., Ong E., Bowden R. Improving visual features for lip-reading. *Proc. of Auditory-Visual Speech Processing (AVSP)*, 2010, pp. 142–147.
17. Khafizov R. G., Yaranceva T. V. Estimation of geometrical distortions of lip contours in visual input systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2017, no. 4, pp. 2–6 (In Russian). doi:10.15217/issn1684-8853.2017.4.2

18. Kukharev G. A., Kamenskaya E. I., Matveev Yu. N., Schegoleva N. L. *Metody obrabotki i raspoznavaniya izobrazhenij lits v zadachakh biometrii* [Methods of Facial Images Processing and Recognition in Biometrics]. Saint-Petersburg, Politekhnika Publ., 2013. 388 p. (In Russian).
19. Ivanko D., Karpov A., Fedotov D., Kipyatkova I., Ryumin D., Ivanko Dm., Minker W., Zelezny M. Multimodal speech recognition: increasing accuracy using high speed video data. *Journal of Multimodal User Interfaces*, 2018, vol. 12, no. 4, pp. 319–328.
20. Ivanko D. V., Kipyatkova I. S., Ronzhin A. L., Karpov A. A. Analysis of multimodal fusion techniques for audio-visual speech recognition. *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki* [Scientific and Technical Journal of Information Technologies, Mechanics and Optics], 2016, vol. 16, no. 3, pp. 387–401 (In Russian).
21. Estellers V., Gurban M., Thiran J. 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.
22. Stewart D., Seymour R., Pass A., Ming J. Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Transactions on Cybernetics*, 2013, vol. 44, no. 2, pp. 175–184.

## Уважаемые авторы!

**При подготовке рукописей статей необходимо руководствоваться следующими рекомендациями.**

Статьи должны содержать изложение новых научных результатов. Название статьи должно быть кратким, но информативным. В названии недопустимо использование сокращений, кроме самых общепринятых (РАН, РФ, САПР и т. п.).

Объем статьи (текст, таблицы, иллюстрации и библиография) не должен превышать эквивалента в 20 страниц, напечатанных на бумаге формата А4 на одной стороне через 1,5 интервала Word шрифтом Times New Roman размером 13, поля не менее двух сантиметров.

Обязательными элементами оформления статьи являются: индекс УДК, заглавие, инициалы и фамилия автора (авторов), ученая степень, звание (при отсутствии — должность), полное название организации, аннотация и ключевые слова на русском и английском языках, ORCID и электронный адрес одного из авторов. При написании аннотации не используйте аббревиатур и не делайте ссылок на источники в списке литературы. Предоставляйте подрисуночные подписи и названия таблиц на русском и английском языках.

Статьи авторов, не имеющих ученой степени, рекомендуется публиковать в соавторстве с научным руководителем, наличие подписи научного руководителя на рукописи обязательно; в случае самостоятельной публикации обязательно предоставляйте заверенную по месту работы рекомендацию научного руководителя с указанием его фамилии, имени, отчества, места работы, должности, ученого звания, ученой степени.

**Формулы** набирайте в Word, не используя формульный редактор (Mathtype или Equation), при необходимости можно использовать формульный редактор; для набора одной формулы не используйте два редактора; при наборе формул в формульном редакторе знаки препинания, ограничивающие формулу, набирайте вместе с формулой; для установки размера шрифта никогда не пользуйтесь вкладкой Other…, используйте заводские установки редактора, не подгоняйте размер символов в формулах под размер шрифта в тексте статьи, не растягивайте и не сжимайте мышью формулы, вставленные в текст; в формулах не отделяйте пробелами знаки: + = –.

Для набора формул в Word никогда не используйте Конструктор (на верхней панели: «Работа с формулами» — «Конструктор»), так как этот ресурс предназначен только для внутреннего использования в Word и не поддерживается программами, предназначенными для изготовления оригинал-макета журнала.

При наборе символов в тексте помните, что символы, обозначаемые латинскими буквами, набираются светлым курсивом, русскими и греческими — светлым прямым, векторы и матрицы — прямым полужирным шрифтом.

**Иллюстрации** предоставляются отдельными исходными файлами, поддающимися редактированию:

— рисунки, графики, диаграммы, блок-схемы предоставляйте в виде отдельных исходных файлов, поддающихся редактированию, используя векторные программы: Visio (*.vsd, *.vsdx); Coreldraw (*.cdr); Excel (*.xls); Word (*.docx); Adobe Illustrator (*.ai); AutoCad (*.dxf); Matlab (*.ps, *.pdf или экспорт в формат *.ai);

— если редактор, в котором Вы изготавливаете рисунок, не позволяет сохранить в векторном формате, используйте функцию экспорта (только по отношению к исходному рисунку), например, в формат *.ai, *.esp, *.wmf, *.emf, *.svg;

— фото и растровые — в формате *.tif, *.png с максимальным разрешением (не менее 300 pixels/inch).

Наличие подрисуночных подписей и названий таблиц на русском и английском языках обязательно (желательно не повторяющих дословно комментарии к рисункам в тексте статьи).

**В редакцию предоставляются:**

— сведения об авторе (фамилия, имя, отчество, место работы, должность, ученое звание, учебное заведение и год его окончания, ученая степень и год защиты диссертации, область научных интересов, количество научных публикаций, домашний и служебный адреса и телефоны, e-mail), фото авторов: анфас, в темной одежде на белом фоне, должны быть видны плечи и грудь, высокая степень четкости изображения без теней и отблесков на лице, фото можно представить в электронном виде в формате *.tif, *.png с максимальным разрешением — не менее 300 pixels/inch при минимальном размере фото 40×55 мм;

— экспертное заключение.

**Список литературы** составляется по порядку ссылок в тексте и оформляется следующим образом:

— для книг и сборников — фамилия и инициалы авторов, полное название книги (сборника), город, издательство, год, общее количество страниц;

— для журнальных статей — фамилия и инициалы авторов, полное название статьи, название журнала, год издания, номер журнала, номера страниц;

— ссылки на иностранную литературу следует давать на языке оригинала без сокращений;

— при использовании web-материалов указывайте адрес сайта и дату обращения.

Список литературы оформляйте двумя отдельными блоками по образцам lit.dot на сайте журнала (http://i-us.ru/paperrules): Литература и References.

Более подробно правила подготовки текста с образцами изложены на нашем сайте в разделе «Правила для авторов».

**Контакты**
Куда: 190000, Санкт-Петербург,
Б. Морская ул., д. 67, ГУАП, РИЦ
Кому: Редакция журнала «Информационно-управляющие системы»
Тел.: (812) 494-70-02
Эл. почта: ius.spb@gmail.com
Сайт: www.i-us.ru