УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2018 Том 18 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2018 Vol. 18 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.93

ПОВЫШЕНИЕ ТОЧНОСТИ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ВИЗУАЛЬНОЙ РУССКОЙ РЕЧИ: ОПТИМИЗАЦИЯ ВИЗЕМНЫХ КЛАССОВ

Д.В. Иванько[a,b,c], Д.В. Федотов[a], А.А. Карпов[b,c]

[a] Ульмский университет, Ульм, 89081, Германия
[b] Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, 199178, Российская Федерация
[c] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: denis.ivanko11@gmail.com

Информация о статье
Поступила в редакцию 29.12.17, принята к печати 30.01.18
doi: 10.17586/2226-1494-2018-18-2-346-349
Язык статьи – русский

Ссылка для цитирования: Иванько Д.В., Федотов Д.В., Карпов А.А. Повышение точности автоматического распознавания визуальной русской речи: оптимизация виземных классов // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 2. С. 346–349. doi: 10.17586/2226-1494-2018-18-2-346-349

Аннотация
В задаче автоматического чтения речи по губам диктора ведутся поиски оптимального набора классов визем, необходимого для максимально эффективного распознавания визуальной речи. Предложен подход для выделения классов визем, позволяющий создавать набор карт соответствия фонема–визема, где каждый класс имеет различное количество визем, от 2 до 48, при неизменном количестве фонем. Виземные классы основаны на их отображении из классов фонем, которые преобразуются в виземные группы в процессе распознавания звучащей речи. Используя полученные карты соответствия, на основе базы данных аудиовизуальной русской речи HAVRUS в работе продемонстрирована зависимость точности распознавания визуальной речи от количества используемых виземных классов. Использование высокоскоростных видеоданных позволило расширить оптимальный набор виземных классов до 20, что привело к улучшению точности распознавания по сравнению с набором из 14 классов.

Ключевые слова
распознавание визуальной речи, виземы, автоматическое чтение речи по губам диктора

Благодарности
Работа выполнена при поддержке Министерства образования и науки Российской Федерации, госзадание № 8.9957.2017/ДААД, а также в рамках бюджетной темы РФ № 0073-2018-0002.

ACCURACY INCREASE FOR AUTOMATIC VISUAL RUSSIAN SPEECH RECOGNITION: VISEME CLASSES OPTIMIZATION

D.V. Ivanko[a,b,c], D.V. Fedotov[a], A. A. Karpov[c,b]

[a] Ulm University, Ulm, 89081, Germany
[b] ITMO University, Saint Petersburg, 197101, Russian Federation
[c] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation
Corresponding author: denis.ivanko11@gmail.com

Article info
Received 29.12.17, accepted 30.01.18
doi: 10.17586/2226-1494-2018-18-2-346-349
Article in Russian
For citation: Ivanko D.V., Fedotov D.V., Karpov A. A. Accuracy increase for automatic visual Russian speech recognition: viseme classes optimization. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 2, pp. 346–349 (in Russian). doi: 10.17586/2226-1494-2018-18-2-346-349

Abstract
Nowadays there are a lot of continuous studies on the correct viseme classes to be used for the most effective automatic lip-reading. The paper proposes a structured approach for the development of speaker-dependent classes of visemes. This method

gives the possibility to create a set of phoneme-viseme correspondence maps, where each class has a different number of visemes from two to forty-eight with a constant number of phonemes. Viseme classes are based on their mapping from phonemes, which are converted into viseme groups during speech recognition process. With the usage of the obtained correspondence maps together with the database of audio-visual Russian speech HAVRUS the paper demonstrates the dependence of the visual speech recognition accuracy on the number of used viseme classes. The application of high-speed video data made it possible to expand the optimal set of viseme classes to twenty that resulted in recognition accuracy improvement by 1.34% compared to the standard set of fourteen classes.

Keywords
visual speech recognition, visemes, automatic lip-reading

Acknowledgements
The research was supported by the Ministry of Education and Science of the Russian Federation, contract No. 8.9957.2017/DAAD, as well as in the framework of the Russian state research No. 0073-2018-0002.

Единственно точного и формализованного определения визем на сегодняшний день не существует, в литературе можно встретить различные варианты [1, 2]. В настоящей работе мы будем использовать следующее определение: виземы – это изображения основных положений губ, соответствующих типовым фонетическим звукам. Таким образом, набор визем по своей природе меньше, чем набор фонем. Это означает, что на каждый класс визем приходится больше обучающих данных (тем самым отчасти устраняется ограничение на объем данных, так как большинство современных баз данных имеют небольшой размер), но в то же время это вводит обобщение между артикуляционными звуками. Для нахождения оптимального количества классов визем необходимо свести к минимуму это обобщение, чтобы максимально повысить точность распознавания речи, а также максимально полно использовать имеющиеся данные.

В качестве базы данных использовался корпус аудиовизуальной русской речи с высокоскоростными видеозаписями HAVRUS [3]. Корпус состоит из записи 20 русских дикторов (10 мужчин и 10 женщин), каждый из которых произносил по 200 подобранных фраз: 130 фраз для обучения были взяты из двух фонетически представительных текстов и были одинаковы для всех дикторов, 70 фраз для тестирования являлись телефонными номерами и отличались для всех дикторов. Видеоданные имеют разрешение 640×480 пикселей и записаны с помощью высокоскоростной камеры с частотой 200 кадров в секунду. Общая длительность аудиоданных приблизительно 6 часов.

Архитектура и принцип работы используемой системы распознавания слитной русской речи более подробно описаны в наших предыдущих работах [4, 5]. В качестве визуальных признаков, описывающих форму губ человека, были использованы пиксельные визуальные признаки на основе метода анализа главных компонент (PCA) [6].

Связь между фонемами (единицами акустической речи) и виземами (единицами визуальной речи) может быть описана с помощью карт соответствия фонема–визема. В [1] показано, как эти карты могут быть получены автоматически из матрицы спутывания фонем. Достоинством этого метода является возможность контролировать, сколько классов визем достаточно для работы системы.

Пример карты соответствия фонема-визема приведен в таблице. В этом случае (20 классам визем соответствуют 48 фонем русского языка) удалось добиться максимальной точности распознавания речи. Дальнейшее увеличение количества виземных классов не приводит к существенному увеличению точности (рисунок). На рисунке приведена зависимость пословной точности распознавания (Word Recognition Rate, WRR) от количества используемых виземных классов. Показаны усредненные значения по 20 дикторам с разбросом, который соответствует значениям лучшей и худшей точности распознавания отдельных дикторов.

Класс виземы	Соответствующие фонемы русской речи	Класс виземы	Соответствующие фонемы русской речи
*V*1	тишина (пауза)	*V*11	/э!/,/э/
*V*2	/а/, /а!/	*V*12	/ы/, /ы!/
*V*3	/и/, /и!/	*V*13	/у/, /у!/
*V*4	/о!/	*V*14	/щ/
*V*5	/б/, /б’/, /п/, /п’/	*V*15	/с/, /с’/,/з/, /з’/,/ц/
*V*6	/ф/, /ф’/, /в/, /в’/	*V*16	/й/
*V*7	/ш/	*V*17	/х/, /х’/
*V*8	/л/, /л’/, /р/, /р’/	*V*18	/ч/
*V*9	/д/, /д’/,/т/, /т’/,/н/, /н’/	*V*19	/м/, /м’/
*V*10	/г/, /г’/, /к/, /к’/	*V*20	/ж/

Таблица. Классы визем и их соответствие фонемам русской речи

Количество используемых виземных классов зависит от языка, и для русского обычно использовалось от 10 до 14 классов [7–9]. В наших экспериментах мы использовали от 2 (разделение на гласные и согласные) до 48 виземных классов (по количеству фонем), с шагом 2. При использовании высокоскоростных видеозаписей (200 кадров в секунду) удается намного лучше отследить быструю динамику движения губ в слитной речи. Исходя из этого, в настоящем исследовании наилучший результат (25,82%) был получен при использовании 20 классов визем, полученная точность распознавания на 1,34% выше, чем при использовании базовых 14 классов (24,48%).

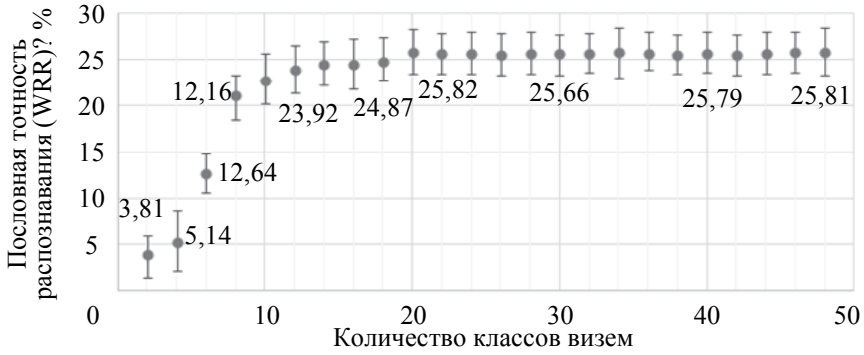

Рисунок. Зависимость пословной точности распознавания русской речи от используемого количества виземных классов

Литература

1. Bear H., Harvey R., Theobald B., Lan Y. Which phoneme-to-viseme maps best improve visual-only computer lip-reading // Lecture Notes in Computer Science. 2014. V. 8888. P. 230–239.
2. Hazen T., Saenko K., La C., Glass J. A segment-based audio-visual speech recognizer: data collection, development, and initial experiments // Proc. 6th Int. Conf. on Multimodal Interfaces. New York, 2004. P. 235–242.
3. Verkhodanova V., Ronzhin A., Kipyatkova I., Ivanko D., Karpov A., Zelezny M. HAVRUS corpus: high-speed recordings of audio-visual Russian speech // Lecture Notes in Computer Science. 2016. V. 9811. P. 338–345.
4. Ivanko D., Karpov A., Ryumin D., Kipyatkova I., Saveliev A., Budkov V., Ivanko Dm., Milos Z. Using a high-speed video camera for robust audio-visual speech recognition in acoustically noisy conditions // Lecture Notes in Computer Science. 2017. V. 10458. P. 757–767.
5. Karpov A. An automatic multimodal speech recognition system with audio and video information // Automation and Remote Control. 2014. V. 75. N 12. P. 2190–2200. doi: 10.1134/S000511791412008X
6. Websdale D., Milner B. Analysing the importance of different visual feature coefficients // Proc. Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing. Vienna, 2015. P. 137–142.
7. Savchenko A., Khokhlova Y. About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems // Optical Memory and Neural Networks. 2014. V. 23. N 1. P. 34–42. doi: 10.3103/S1060992X14010068
8. Zheng G.L., Zhu M., Feng L. Review of lip-reading recognition // Proc. 7th International Symposium on Computational Intelligence and Design. Hangzhou, China, 2014. P. 293–298. doi: 10.1109/ISCID.2014.110
9. 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. V. 8773. P. 50–57.

References

1. Bear H., Harvey R., Theobald B., Lan Y. Which phoneme-to-viseme maps best improve visual-only computer lip-readingю *Lecture Notes in Computer Science*, 2014, vol. 8888, pp. 230–239.
2. Hazen T., Saenko K., La C., Glass J. A segment-based audio-visual speech recognizer: data collection, development, and initial experiments. *Proc. 6th Int. Conf. on Multimodal Interfaces*. New York, 2004, pp. 235–242.
3. Verkhodanova V., Ronzhin A., Kipyatkova I., Ivanko D., Karpov A., Zelezny M. HAVRUS corpus: high-speed recordings of audio-visual Russian speech. *Lecture Notes in Computer Science*, 2016, vol. 9811, pp. 338–345.
4. Ivanko D., Karpov A., Ryumin D., Kipyatkova I., Saveliev A., Budkov V., Ivanko Dm., Milos Z. Using a high-speed video camera for robust audio-visual speech recognition in acoustically noisy conditions. *Lecture Notes in Computer Science*, 2017, vol. 10458, pp. 757–767.
5. Karpov 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
6. Websdale D., Milner B. Analysing the importance of different visual feature coefficients. *Proc. Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing*. Vienna, 2015, pp. 137–142.
7. Savchenko A., Khokhlova Y. About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems. *Optical Memory and Neural Networks*, 2014, vol. 23, no. 1, pp. 34–42. doi: 10.3103/S1060992X14010068
8. Zheng G.L., Zhu M., Feng L. Review of lip-reading recognition. *Proc. 7th International Symposium on Computational Intelligence and Design*. Hangzhou, China, 2014, pp. 293–298. doi: 10.1109/ISCID.2014.110
9. 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.

Авторы

Иванько Денис Викторович – аспирант, Ульмский университет, Ульм, 89081, Германия; младший научный сотрудник, Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург,

Authors

Denis V. Ivanko – postgraduate, Ulm University, Ulm, 89081, Germany; junior scientific researcher, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), Saint Petersburg, 199178, Russian Federation;

199178, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57190967993, ORCID ID: 0000-0003-0412-7765, denis.ivanko11@gmail.com

Федотов Дмитрий Валерьевич – аспирант, Ульмский университет, Ульм, 89081, Германия, Scopus ID: 57195680712, ORCID ID:: 0000-0001-5401-588X, dmitrii.fedotov@uni-ulm.de

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; Scopus ID: 57195330987, ORCID ID:: 0000-0003-3424-652X, karpov@iias.spb.su

postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 57190967993, ORCID ID: 0000-0003-0412-7765, denis.ivanko11@gmail.com

Dmitry V. Fedotov – postgraduate, Ulm University, Ulm, 89081, Germany, Scopus ID: 57195680712, ORCID ID: 0000-0001-5401-588X, dmitrii.fedotov@uni-ulm.de

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; Scopus ID: 57195330987, ORCID ID: 0000-0003-3424-652X, karpov@iias.spb.su