Автоматическое определение количества минимальных единиц языка по артикуляции

В.О. Ячная [1,2], В.Р. Луцив [1]
[1] Государственный университет аэрокосмического приборостроения, 190000, Россия, г. Санкт-Петербург, ул. Большая Морская, д. 67, лит. а,
[2] Институт физиологии имени И.П. Павлова РАН, 199034, Россия, Санкт-Петербург, наб. Макарова, д. 6

Аннотация

Представленная работа посвящена автоматическому анализу паравербального компонента общения человека. В статье описаны системы, определяющие количество минимальных языковых единиц (слогов и фонем) в устной речи по видеоданным. Такие системы могут быть использованы в оценке темпа артикулирования говорящего, что может применяться в доклинической диагностике некоторых патологических состояний или определении эмоционального статуса. Для проведения исследования была модифицирована существующая база данных слов английского языка и получена разметка, содержащая информацию о количестве слогов и фонем в каждом слове. В ходе исследования адаптирована система распознавания слов для решения поставленной задачи, а также разработана новая архитектура нейронной сети для определения количества слогов и фонем в слове. Оценка эффективности разработанных систем производилась как на наборах заранее известных системам слов, так и на новых для них словах. В результате работы получена система, определяющая количество минимальных единиц языка в произнесённом слове, предоставляющая возможность последующей оценки темпа артикулирования информанта.

Ключевые слова: распознавание речи, артикуляция, компьютерное зрение, нейронные сети.

Цитирование: **Ячная, В.О.** Автоматическое определение количества минимальных единиц языка по артикуляции / В.О. Ячная, В.Р. Луцив // Компьютерная оптика. – 2024. – Т. 48, № 6. – С. 956-962. – DOI: 10.18287/2412-6179-CO-1451.

Citation: Yachnaya VO, Lutsiv VR. Automatic estimation of the number of minimal language units by articulation. Computer Optics 2024; 48(6): 956-962. DOI: 10.18287/2412-6179-CO-1451.

Введение

Задача распознавания устной речи представляет собой процесс преобразования акустических и визуальных сигналов в текстовое представление. Технология распознавания речи на основе аудиоданных уже используется на практике [1], а в последние годы развивается и технология чтения по губам (распознавания артикуляции по видеоданным) [2]. Использование видеоданных обуславливается рядом причин, среди которых использование систем в условиях, затрудняющих передачу аудиосигнала, или при нарушениях речевого аппарата человека [1, 2]. Более подробно необходимость анализа видеоданных, а также возможные приложения описываются в [2], где также отмечается, что в настоящий момент наиболее популярными и эффективными методами анализа являются методы, основанные на применении искусственных нейронных сетей (ИНС).

Распознавание устной речи по видеоданным производится обычно на уровне слов [2], но также может производиться и на уровне более мелких единиц языка, например, фонем [3 – 5]. Исходя из этого формируются смежные задачи, посвящённые анализу таких языковых единиц. Это могут быть, например, задачи отображения фонемы в визему и наоборот [6 – 9], отображения слогов в фонемы [10] и иные [11].

Исследования, посвященные минимальным единицам языка, находят своё применение на практике. Так, анализ количества таких мелких языковых единиц (как и слов целиком [12]) в единицу времени или в лексической конструкции называется темпом артикулирования и может быть использован в диагностике некоторых патологических состояний человека, связанных с нарушениями речи (например, фонологическое расстройство [13]) или с изменениями когнитивных способностей пациента (как болезнь Альцгеймера [14]), а также при определении эмоционального статуса человека [15, 16] (в частности, влияние фрустрации [17]). Подобные исследования могут позволить автоматически анализировать сразу два компонента общения: вербальный, отвечающий за информационную составляющую передаваемого сообщения, и паравербальный (например, темп речи), вносящий в устную речь дополнительные значения [2].

Кроме того, количество произнесённых единиц языка важно в задачах определения ошибок в искусственно сгенерированных данных. Например, работа [18] посвящена определению ошибок в синтетических видео по выявленным несоответствиям фонем и визем.

Отдельно отметим, что в большинстве исследований выделяются такие единицы языка, как фонемы или слоги, однако в зависимости от исследуемого естественного языка могут использоваться и другие (например, мора в японском языке в [1]).

В связи с вышеописанной важностью анализа языковых единиц в речи человека, целью данной работы является разработка системы на основе искусственной нейронной сети, определяющей количество слогов и количество фонем в произнесённом слове. В ходе выполнения работы на основе базы данных (БД) Lip Reading in the Wild (LRW) [19] произведена разметка, описывающая слоговый и фонемный состав слов данной базы. Для решения поставленной задачи далее было выделено три подхода:

1. Адаптация системы распознавания слов по артикуляции для определения количества слогов и фонем в словах, на которых была обучена система.
2. Разработка новой системы определения количества слогов и фонем по артикуляции на известных системе словах.
3. Оценка эффективности обеих систем для определения количества слогов и фонем в словах, не известных системам.

В первых двух случаях исследуется работа каждой из систем на словах, примеры которых присутствовали в обучающей выборке БД. В третьем случае исследуется поведение обеих описанных систем на словах, примеры которых не были представлены в обучающей выборке, тем самым имитируется практическое применение систем. Дополнительно рассматриваются некоторые возможные модификации разрабатываемой системы определения количества слогов и фонем в словах.

1. Подготовка данных

В качестве основы для работы была выбрана БД LRW [19]. БД LRW содержит 500 слов (классов) английского языка, произнесённых сотнями разных информантов. Для каждого из слов представлено 800 – 1000 обучающих примеров, 50 валидационных примеров и 50 тестовых примеров. Каждый пример представляет собой видеофрагмент из новостной телепередачи продолжительностью 1,16 секунды, а длительность слова указана в метаданных, по которым можно определить начальный и конечный кадры.

На основе БД LRW были созданы две новые БД, содержащие информацию о количестве в слове слогов и фонем: LRW-Syl (от англ. syllable) и LRW-Ph (от англ. phoneme). БД LRW-Syl содержит видео с примерами произнесения 500 различных слов английского языка из БД LRW и разметку, описывающую количество слогов в каждом слове. Аналогично, БД LRW-Ph содержит все примеры из БД LRW с разметкой, указывающей количество фонем в каждом примере. При этом примеры, в исходной БД LRW входившие в обучающую выборку, в БД LRW-Syl и LRW-Ph также попали в обучающую выборку, и аналогично с примерами валидационной и тестовой выборок. Отметим, что при таком разбиении данных примеры каждого из 500 слов попадают и в обучающую, и в валидационную, и в тестовую выборки.

Параметры полученных наборов данных приведены в табл. 1 и 2. Базы LRW-Syl и LRW-Ph используются в дальнейших экспериментах по определению количества слогов и фонем в слове.

Количество слогов в слове определялось с помощью ресурса [20], а количество фонем – с помощью словаря The Carnegie Mellon University (CMU) Pronouncing Dictionary [21]. Словарь CMU [21] содержит 39 фонем для североамериканского варианта английского языка. Согласно этим ресурсам, например, слово “ABSOLUTELY” содержит 4 слога (“AB-SO-LUTE-LY”) и 9 фонем (“AE-B-S-AH-L-UW-T-L-IY”).

Табл. 1. Параметры БД LRW-Syl

Количество		**Примеры, тыс. шт.**		
Слоги	**Слова**	**Обучающие**	**Валидационные**	**Тестовые**
1	125	122,39	6,25	6,25
2	260	254,99	13,00	13,00
3	86	83,15	4,30	4,30
4	29	28,24	1,45	1,45

Табл. 2. Параметры БД LRW-Ph

Количество		**Примеры, тыс. шт.**		
Фонемы	**Слова**	**Обучающие**	**Валидационные**	**Тестовые**
3	38	37,52	1,90	1,90
4	116	113,54	5,80	5,80
5	128	124,80	6,40	6,40
6	97	95,89	4,85	4,85
7	59	57,22	2,95	2,95
8	40	38,26	2,00	2,00
9	15	14,90	0,75	0,75
10	5	4,70	0,25	0,25
11	2	1,94	0,10	0,10

Анализ слов, представленных в БД LRW, с точки зрения количества фонем и слогов приводится в [22].

2. Моделирование и применение системы на известном наборе слов

2.1. Адаптация системы распознавания слов для определения количества слогов и фонем в словах

В данном параграфе изучается работа системы распознавания слов по артикуляции человека для определения количества слогов и фонем в произносимом слове. Поскольку количество фонем и слогов в известном слове заведомо известно (например, в соответствии с описанными ранее словарями), то система, корректно распознающая слова, вместе с этим так же корректно определяет количество фонем и слогов в произнесённом слове.

В качестве инструмента для определения слов была выбрана система на основе ИНС из работы [23]. Данная модель, как и большинство моделей глубоко-

го обучения для распознавания речи по артикуляции [2], имеет двухмодульную структуру. Внешний модуль (frontend) для выделения признаков использует свёрточную нейронную сеть на основе ResNet-18 [24] с трёхмерной свёрткой (3-Dimension Convolution – C3D) с размером ядра 5×7×7 [25]. Архитектура ResNet представляет собой универсальное средство, зарекомендовавшее себя как эффективный инструмент для распознавания артикуляции (например, в работах [26 – 29]). Во внутреннем модуле (backend) для моделирования долгосрочных зависимостей применяется рекуррентная нейронная сеть – BiGRU, представляющая собой двунаправленную версию архитектуры Управляемых Рекуррентных Нейронов (Bidirectional Gated Recurrent Units – GRU). В качестве входных данных модель принимает видеоролик, содержащий пример произнесения одного слова.

Данная модель была обучена авторами распознавать 500 слов на английском языке из БД LRW [19] с показателем точности 85 % на тестовой выборке. Соответственно, эта модель [23] с точностью не менее 85 % способна определить количество фонем и слогов в словах из БД LRW.

Нами были подробно рассмотрены результаты работы данной системы. Мы применили предоставленную модель к видеороликам, содержащимся в тестовой выборке БД LRW, и получили заявленные 85 % точности определения слова. Далее мы проанализировали те 15 % видеофрагментов, которые модель некорректно распознала, исходя из того, что система могла неправильно определить слово, однако принять его за слово, содержащее то же количество слогов или фонем.

Например, на одном из видеороликов система приняла содержащееся в нём слово “ABOUT” за слово “BLACK”. Слово “ABOUT” состоит из 2 слогов и 4 фонем, тогда как слово “BLACK” – из 1 слога и 4 фонем. В данном случае система правильно определила только количество фонем.

Таким образом, если оценивать модель [23] с точки зрения распознавания количества слогов и количества фонем, то в соответствии со словарями [20] и [21] её точность на тестовой выборке БД LRW составляет уже не 85 %, а 92,812 % для слогов и 89,24 % для фонем. Матрицы неточности этой модели и их описание представлены в параграфе 4.

2.2. Разработка системы определения количества фонем и слогов в словах

Далее нами была построена и обучена система на основе ИНС для определения количества слогов и фонем в слове по артикуляции. Дополнительно были проанализированы некоторые возможные модификации этой архитектуры.

За основу мы взяли архитектуру, описанную в [23], и адаптировали под решение поставленной задачи. Наша ИНС также состоит из двух модулей: архитектуры C3D-ResNet-18 с размером ядра 5×7×7 во внешнем и BiGRU во внутреннем. Размер выходного слоя зависит от количества классов: 4 для определения количества слогов и 9 для определения количества фонем.

Описанная нейронная сеть далее в работе обозначается как C3D-ResNet18. Данная архитектура используется в определении количества слогов и фонем, результаты практического моделирования которого описаны ниже.

Для обучения архитектуры использовалась функция потерь перекрестной энтропии с коэффициентом скорости обучения 0,00001, к которому применялся планировщик Косинусный отжиг, сбрасывающий темп обучения и действующий как симулированный перезапуск процесса обучения, а в качестве оптимизатора применялся метод Адаптивной оценки момента (Adam).

Определение количества слогов. В результате обучения нейронной сети C3D-ResNet18 на БД LRW-Syl была получена точность определения количества слогов на тестовой выборке 90,104 %. На рис. 1*б* представлена матрица неточности полученной модели. Из неё видно, что несмотря на то, что БД имеет неравномерность распределения обучающих примеров между классами, смещения предсказаний сети не наблюдается. Например, класс «4», представленный наименьшим количеством видеофрагментов, распознаётся моделью с точностью 89 %, тогда точность распознавания класса «3», обучающих примеров которого в 3 раза больше, меньше – 86 %.

Определение количества фонем. Аналогично, сеть C3D-ResNet18 была обучена на БД LRW-Ph для определения количества фонем. Несмотря на то, что эта БД также несбалансирована (как и в случае БД LRW-Syl), обученная нейронная сеть достигает точности распознавания на тестовой выборке – 86,948 %.

Блок Squeeze-and-Excitation. В смежной работе [23] во внешнем модуле системы используется дополнительный блок Сжатия-и-Стимуляции (Squeeze-and-Excitation – SE) [30]. Данный блок реализует механизм взвешивания каналов свёрточных блоков, чтобы нейронная сеть могла адаптивно регулировать вес каждой карты признаков. Как отмечается авторами [23], блок SE повышает точность определения слов по артикуляции [23 – 24]. В связи с этим нами были проведены дополнительные эксперименты и оценено влияние блока SE на точность работы разрабатываемой системы. Такая архитектура далее обозначается как SE-C3D-ResNet18, результаты её обучения приведены в параграфе 4. Исходя из полученных значений при использовании модуля SE точность на тестовой выборке снижается в среднем на 0,87 %. Поскольку для решения поставленной задачи применение блока SE оказывается неэффективным, он не использовался в дальнейших экспериментах.

Увеличение глубины сети. В настоящей работе нами также проведено исследование, оценивающее влияние глубины нейронной сети в основе внешнего модуля системы на точность определения количества слогов и фонем в слове. А именно, сравнивается описанная базовая модель C3D-ResNet18 и модель C3D-ResNet34.

Результаты обучения модели C3D-ResNet34 приведены в параграфе 4. Из полученных результатов видно, что использование более глубокой нейронной сети ResNet-34 в качестве внешнего модуля не даёт прироста точности.

Полученные результаты согласуются с проведённым в [23] анализом, где менее глубокая сеть ResNet-18 на базе данных CAS-VSR-W1k (LRW-1000) [31] показала более высокие результаты по распознаванию слов, чем более глубокая сеть ResNet-34, в связи с чем в качестве итоговой модели авторами используется ResNet-18. Кроме того, в ряде работ ([26–28]) предпочтение также отдаётся менее глубокой архитектуре. Исходя из этого нами далее рассматривается архитектура ResNet-18 в качестве внешнего модуля системы.

3. Применение систем на наборе неизвестных слов

Анализируемый на практике видеопоток может содержать слова, которые не входили в состав обучающей выборки. Поскольку полученные в предыдущем параграфе результаты отражают точности, достигаемые на словах, примеры которых входили в обучающую выборку, нами дополнительно были проанализированы следующие ситуации:

1. Оценка точности определения количества слогов и фонем моделью, обученной для распознавания слов на неизвестных системе словах.
2. Оценка точности работы разработанной архитектуры C3D-ResNet18 для определения количества слогов и фонем на неизвестных системе словах.

Для выполнения подобного моделирования были перераспределены примеры из БД LRW следующим образом: для каждого из классов (для 4 в случае определения количества слогов и 11 – фонем) 80 % слов попали в обучающую выборку и по 10 % слов – в валидационную и тестовую. Таким образом, все видеопримеры каждого слова оказались или в обучающей, или в валидационной, или в тестовой выборке. Поскольку классы «10» и «11» в БД LRW-Ph представлены слишком малым количеством различных слов для подобного разбиения (табл. 2), то далее они исключаются из распознавания.

Полученные наборы данных далее обозначаются LRW-Syl-W и LRW-Ph-W, где «W» отмечает произведённое нами разбиение по словам (от англ. word), а их подробные параметры приведены в табл. 3 и 4.

Использование системы для распознавания слов. В ходе моделирования нами была обучена система распознавания слов из работы [23] на 80 % слов из БД LRW. При этом эти 80 % обучающих слов были определены не случайным образом, а взяты из обучающих выборок наборов LRW-Syl-W и LRW-Ph-W в соответствии с решаемой задачей.

Табл. 3. Параметры БД LRW-Syl-W

Кол-во слогов	Кол-во слов	Кол-во обучающих слов	Кол-во валидационных слов	Кол-во тестовых слов
1	125	100	13	12
2	260	208	26	26
3	86	69	9	8
4	29	24	3	2

Табл. 4. Параметры БД LRW-Ph-W

Кол-во фонем	Кол-во слов	Кол-во обучающих слов	Кол-во валидационных слов	Кол-во тестовых слов
3	38	31	4	3
4	116	93	12	11
5	128	103	13	12
6	97	78	10	9
7	59	48	6	5
8	40	32	4	4
9	15	12	2	1

При использовании описанной методики точность определения количества слогов составила 40,56 %, а количества фонем – 19,39 %, что незначительно превышает вероятность случайного верного определения (25 % и 14,28 % соответственно). Данные результаты приведены в табл. 5.

Использование системы для определения количества слогов и фонем. В результате обучения архитектуры C3D-ResNet18 на БД LRW-Syl-W и LRW-Ph-W точность определения количества слогов в неизвестных словах составляет 58,667 %, фонем – 31,69 %. Данные результаты приведены в табл. 5.

4. Результаты и дискуссия

Ниже приведена табл. 5, содержащая результаты всех описанных экспериментов, а также рис. 1 – 4, иллюстрирующие матрицы неточности проанализированных моделей.

Табл. 5. Точность определения количества слогов и фонем описанными методами

База данных	Архитектура	Точность на тестовой выборке, %
LRW-Syl	[23]	**92,812**
	C3D-ResNet18	90,104
	SE-C3D-ResNet18	89,930
	C3D-ResNet34	89,980
LRW-Ph	[23]	**89,240**
	C3D-ResNet18	86,948
	SE-C3D-ResNet18	85,376
	C3D-ResNet34	84,828
LRW-Syl-W	[23]	40,560
	C3D-ResNet18	**58,667**
LRW-Ph-W	[23]	19,390
	C3D-ResNet18	**31,690**

Несмотря на то, что в обеих БД LRW-Syl и LRW-Ph резко отличается количество обучающих примеров в разных классах, согласно матрицам неточности, показанным на рис. 1 и рис. 3, нет прямой зависимости между количеством обучающих примеров в классе и точностью распознавания этого класса. Например, согласно рис. 3 наибольшую точность распознавания имеет класс «10», представленный малым количеством обучающих примеров. Однако в случае использования моделей на наборах неизвестных для неё слов заметно увеличение частоты предсказания сети тех классов, которые в обучающей выборке представлены наибольшим количеством примеров (рис. 2 и рис. 4).

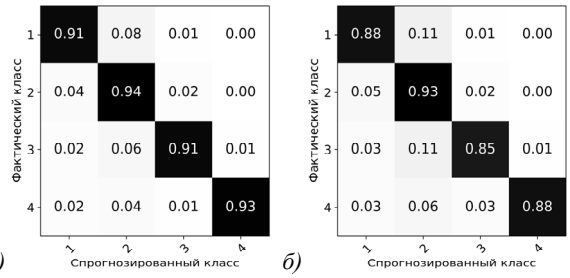

Рис. 1. Матрицы неточности систем определения количества слогов на БД LRW-Syl. (а) архитектура [23], (б) C3D-ResNet18

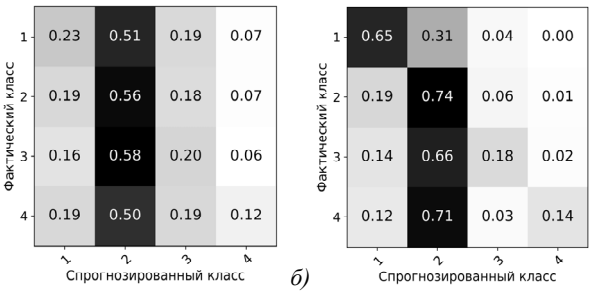

Рис. 2. Матрицы неточности систем определения количества слогов на новых словах (БД LRW-Syl-W). (а) архитектура [23], (б) C3D-ResNet18

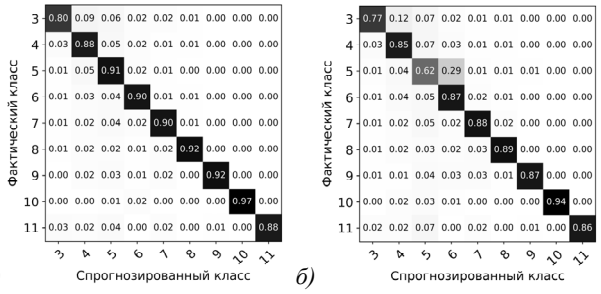

Рис. 3. Матрицы неточности систем определения количества фонем на БД LRW-Ph. (а) архитектура [23], (б) C3D-ResNet18

Исходя из полученных результатов (табл. 5) видно, что базовая архитектура C3D-ResNet18 достигает наибольшей точности определения количества слогов и фонем в слове. Ни добавление блока SE, предлагаемого в [23], ни использование более глубокой архитектуры ResNet-34 не повышают точность на тестовой выборке для решения поставленных в работе задач.

С точки зрения определения количества слогов и фонем в слове разработанная в данной статье архитектура C3D-ResNet18 уступает системе [23], предназначенной для определения слов, только в случае анализа тех слов, которые использовались при обучении моделей. Однако при анализе неизвестных системам слов разработанная архитектура C3D-ResNet18 показывает лучшие результаты. В целом невысокая точность систем (до 58,667 %) может объясняться несбалансированностью обучающей выборки и последующим смещением предсказаний нейронной сети, а также возможным переобучением алгоритмов на словах. В таком случае путём решения может послужить увеличение обучающей выборки, в том числе добавлением примеров слов, не входящих в БД LRW.

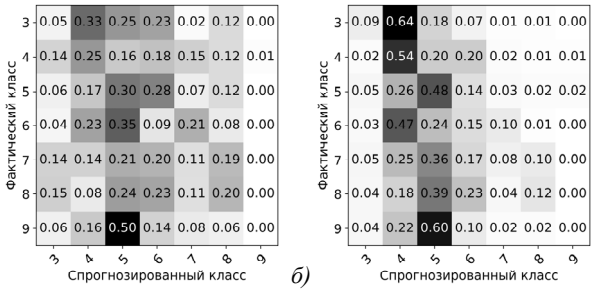

Рис. 4. Матрицы неточности модели систем определения количества фонем на новых словах (БД LRW-Ph-W). (а) архитектура [23], (б) C3D-ResNet18

Заключение

Данная работа освещает проблему определения количества слогов и фонем в слове. В частности, в статье описываются системы, предназначенные для определения количества слогов и фонем в слове английского языка по артикуляции человека.

Для решения поставленной задачи были составлены две новые БД LRW-Syl и LRW-Ph, содержащие информацию о количестве слогов и фонем слов английского языка, представленных в БД LRW [19].

В результате работы получена архитектура нейронной сети для определения количества слогов и фонем в слове, основанная на C3D-ResNet18 для внешнего модуля и BiGRU для внутреннего модуля, а также проведён сравнительный анализ её эффективности по отношению к работе системы распознавания слов. В ходе экспериментов установлено, что при анализе слов, примеры которых присутствовали в обучающей выборке, рассмотренные системы сравнимы по точности работы (однако наибольшей точности определения количества слогов и фонем достигает система, предназначенная для распознавания слов). В то же время при анализе неизвестных заранее слов система, спроектированная специально для определения количества слогов и фонем, оказывается эффективнее. Полученные в этом случае показатели, однако, недостаточны для практического применения.

Тем не менее стоит отметить определённый прогресс в развитии систем распознавания в обсуждаемой области по сравнению с известными на данный момент аналогами. Результаты могут быть также основанием для расширения обучающей выборки с точки зрения лексического разнообразия разговорной речи, а также модификаций и последующего дообучения рассмотренных систем. Так, несмотря на то, что на данный момент уже разработаны эффективные системы распознавания слов по артикуляции, решение задачи определения количества малых единиц языка является достаточно сложным. Кроме того, установлены ограничения обобщающей способности системы распознавания слов применительно к поставленной задаче. Также продемонстрированы подходы к адаптации системы к конкретной задаче, позволяющие повысить точность её решения.

В связи с актуальностью рассмотренной проблемы представляются интересными дальнейшие исследования, которые могут способствовать одновременному анализу вербального и паравербального компонентов общения. Поскольку рассматривался такой паравербальный компонент общения, как темп артикуляции, то такими направлениями могут быть доклиническая диагностика патологических состояний или функциональных нарушений, связанных с изменениями лексического состава устной речи, а также определение эмоционального статуса информанта без привлечения соответствующих экспертов в ходе проведения различных опросов или других социальных исследований.

Система, подобная разработанной, также может быть использована для улучшения показателей работоспособности мультимодальной системы распознавания устной речи в качестве модуля, дополняющего или подтверждающего результаты из параллельного канала данных.

Благодарности

Работа поддержана средствами федерального бюджета в рамках государственного задания ФГБУН Институт физиологии им. И.П. Павлова РАН (№ 1021062411653-4-3.1.8).

References

[1] Arakane T, Saitoh T, Chiba R, Morise M, Oda Y. Conformer-based lip-reading for Japanese sentence. In Book: Yan WQ, Nguyen M, Stommel M, eds. Image and Vision Computing. Cham: Springer; 2023. DOI: 10.1007/978-3-031-25825-1_34.

[2] Yachnaya VO, Lutsiv VR, Malashin RO. Modern automatic recognition technologies for visual communication tools. Computer Optics 2023; 47(2): 287-305. DOI: 10.18287/2412-6179-CO-1154.

[3] Yu C, Yu J, Qian Z, Tan Y. Improvement of acoustic models fused with lip visual information for low-resource speech. Sensors 2023; 23(4): 2071. DOI: 10.3390/s23042071.

[4] El-Bialy R, et al. Developing phoneme-based lip-reading sentences system for silent speech recognition. CAAI Trans Intell Technol 2023; 8(1): 129-138. DOI: 10.1049/cit2.12131.

[5] Rahmani MH, Almasganj F. Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features. 3rd Int Conf on Pattern Recognition and Image Analysis (IPRIA) 2017: 195-199. DOI: 10.1109/PRIA.2017.7983045.

[6] Ivanko D, Ryumin D, Kipyatkova I, Axyonov A, Karpov A. Lip-reading using pixel-based and geometry-based features for multimodal human–robot interfaces. Proc 14th Int Conf on Electromechanics and Robotics “Zavalishin's Readings”. Smart Innovation, Systems and Technologies 2020: 154. DOI: 10.1007/978-981-13-9267-2_39.

[7] Fernandez-Lopez A, Sukno FM. Optimizing phoneme-to-viseme mapping for continuous lip-reading in Spanish. In Book: Cláudio AP, Bechmann D, Richard P, Yamaguchi T, Linsen L, Telea A, Imai F, Tremeau A, eds. Computer Vision, Imaging and Computer Graphics – Theory and Applications. Cham: Springer; 2019: 305-328. DOI: 10.1007/978-3-030-12209-6_15.

[8] Rachman A, Hidayat R, Nugroho H. Improving phoneme to viseme mapping for indonesian language. Int J Inform Technol Electrical Eng 2020; 4(1): 1-7. DOI: 10.22146/ijitee.47577.

[9] Wakkumbura WGVK, Madhubhashana RAH, Alahakoon PMK, Kumara WGCW, Hinas MNA. Phoneme-viseme mapping for sinhala speaking robot for Sri Lankan healthcare applications. IEEE 4th Eurasia Conf on Biomedical Engineering, Healthcare and Sustainability (ECBIOS) 2022; 258-262. DOI: 10.1109/ECBIOS54627.2022.9945003.

[10] Srivastava T, Khanna P, Pan S, Nguyen P, Jain S. MuteIt: Jaw motion based unvoiced command recognition using earable. Proc ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2022; 6(3): 140. DOI: 10.1145/3550281.

[11] Lee CH, Kwon Y, Kim K. Syllable transposition effects in Korean word recognition. J Psycholinguist Res 2015; 44: 309-315. DOI: 10.1007/s10936-015-9353-7.

[12] Diaz-Asper M, Holmlund TB, Chandler C, Diaz-Asper C, Foltz PW, Cohen AS, Elvevåg B. Using automated syllable counting to detect missing information in speech transcripts from clinical settings. Psychiatry Res 2022; 315: 114712. DOI: 10.1016/j.psychres.2022.114712.

[13] Wertzner HF, Silva LM. Speech rate in children with and without phonological disorder. Pró-Fono Revista de Atualização Científica 2009; 21(1): 19-24. DOI: 10.1590/S0104-56872009000100004.

[14] Brewer E, Mirheidari B, O'Malley R, Reuber M, Christensen H, Blackburn DJ. Characterising spoken interactions of healthy ageing adults with CognoSpeak, a web-based cognitive assessment tool. Alzheimer's Dementia 2021; 17: e052913. DOI: 10.1002/alz.052913.

[15] Isaeva AA. Influence of emotional tension on speech production [In Russian]. Proceedings of VSU Series: Linguistics and Intercultural Communication 2023; 4: 34-41. DOI: 10.17308/lic/1680-5755/2022/4/34-41.

[16] Horkous H, Mhania G. Speech emotions recognition of joy and sadness based on prosodic and mfccs parameters. Models & Optimisation and Mathematical Analysis Journal 2018; 6(1): 15-18. Source: <https://www.asjp.cerist.dz/en/article/71119>.

[17] Kuznetsova YM, Kuruzov IA, Smirnov IV, Stankevich MA, Starostina EV, Chudova NV. Textual manifestations of social

network user frustration [In Russian]. Media Linguistics 2020; 7(1): 4-15. DOI: 10.21638/spbu22.2020.101.
[18] Agarwal S, Farid H, Fried O, Agrawala M. Detecting deep-fake videos from phoneme-viseme mismatches. IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops (CVPRW) 2020: 2814-2822. DOI: 10.1109/CVPRW50498.2020.00338.
[19] The Oxford-BBC Lip Reading in the Wild (LRW) Dataset. Source: <https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html>.
[20] How Many Syllables. Source: <https://www.howmanysyllables.com>.
[21] The Carnegie Mellon University (CMU) pronouncing dictionary. Source: <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
[22] Yachnaya VO. The role of the number of smallest linguistic units in a word on the accuracy of the word-level lip-reading system. 8th Int Conf "Video and Audio Signal Processing in the Context of Neurotechnologies" (SPCN-2023) 2023.
[23] Feng D, Yang S, Shan S. An efficient software for building LIP reading models without pains. IEEE Int Conf on Multimedia & Expo Workshops (ICMEW) 2021: 1-2. DOI: 10.1109/ICMEW53276.2021.9456014.
[24] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2016: 770-778. DOI: 10.1109/CVPR.2016.90.
[25] Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading. Proc Interspeech 2017: 3652-3656. DOI: 10.21437/Interspeech.2017-85.
[26] Ma P, Martinez B, Petridis S, Pantic M. Towards practical lipreading with distilled and efficient models. 2021-2021 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2021: 7608-7612. DOI: 10.1109/ICASSP39728.2021.9415063.
[27] Arakane T, Saitoh T. Efficient DNN model for word lip-reading. Algorithms 2023; 16(6): 269. DOI: 10.3390/a16060269.
[28] Tsourounis D, Kastaniotis D, Fotopoulos S. Lip reading by alternating between spatiotemporal and spatial convolutions. J Imaging 2021; 7(5): 91. DOI: 10.3390/jimaging7050091.
[29] Naif KS, Hashim, Kadhim Mahdi. Automatic lip reading for decimal digits using ResNet50 Model. Journal of College of Education for Pure Science 2022; 12(2): 308. DOI: 10.32792/utq.jceps.12.02.30.
[30] Hu J, Shen L, Sun G. Squeeze and excitation networks. IEEE/CVF Conf on Computer Vision and Pattern Recognition 2018: 7132-7141. DOI: 10.1109/CVPR.2018.00745.
[31] Yang S, Zhang Y, Feng D, Yang M, Wang C, Xiao J, Long K, Shan S, Chen X. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. 14th IEEE Int Conf on Automatic Face & Gesture Recognition (FG 2019) 2019: 1-8. DOI: 10.1109/FG.2019.8756582.

Сведения об авторах

Ячная Валерия Олеговна, 1997 года рождения, в 2021 году окончила Санкт-Петербургский государственный университет аэрокосмического приборостроения по специальности 09.04.01 «Информатика и вычислительная техника». Проходит обучение по программе аспирантуры в Санкт-Петербургском государственном университете аэрокосмического приборостроения по специальности 09.06.01 «Информатика и вычислительная техника». Работает младшим научным сотрудником в Институте физиологии им. И.П. Павлова РАН. Область научных интересов: компьютерное зрение. E-mail: *tamimio.yvo@hotmail.com*

Луцив Вадим Ростиславович, 1954 года рождения, в 1977 году окончил Ленинградский институт авиационного приборостроения по специальности 0608 «Электронные вычислительные машины», в 2012 году решением диссертационного совета при Санкт-Петербургском государственном университете аэрокосмического приборостроения (ГУАП) присуждена ученая степень доктора технических наук, профессор ГУАП.
E-mail: *vluciv@mail.ru*

ГРНТИ: 28.23.15
Поступила в редакцию 3 ноября 2023 г. Окончательный вариант – 13 марта 2024 г.

Automatic estimation of the number of minimal language units by articulation

V.O. Yachnaya [1,2], V.R. Lutsiv [1]
[1] Saint Petersburg State University of Aerospace Instrumentation,
190000, Saint-Petersburg, Russia, Bolshaya Morskaya 67,
[2] Pavlov Institute of Physiology, Russian Academy of Sciences,
199034, Saint-Petersburg, Russia, Naberezhnaya Makarova 6

Abstract

The presented work is dedicated to the automatic analysis of the paraverbal component of human communication. The article describes systems that determine the number of minimal linguistic units (syllables and phonemes) in spoken language based on video data. Such systems can be used to assess the subject speech rate, which can be applied in the preclinical diagnosis of certain pathological conditions or determining emotional status. To conduct the research, an existing database of English words was modified, and annotations containing information on the number of syllables and phonemes in each word were obtained. During the study, a word recognition system was adapted to solve the stated task, and a new neural network architecture to determine the number of syllables and phonemes in a word was designed. The effectiveness of the developed systems was assessed on both sets of previously known to the systems words and on new words. As a result of the research, a system that determines the number of minimal language units in a spoken word was obtained, providing the opportunity for subsequent assessment of the subject articulation rate.

Keywords: visual speech recognition, articulation, computer vision, neural networks.

Citation: Yachnaya VO, Lutsiv VR. Automatic estimation of the number of minimal language units by articulation. Computer Optics 2024; 48(6): 956-962. DOI: 10.18287/2412-6179-CO-1451.

Acknowledgements: This study was supported by the State Program 47 GP "Scientific and Technological Development of the Russian Federation "(2019-2030), theme 0134-2019-0006.

Authors' information

Valeriya Olegovna Yachnaya (b. 1997) graduated from Saint Petersburg State University of Aerospace Instrumentation in 2021, majoring in Computer Science and Engineering. Currently she is a graduate student at Saint Petersburg State University of Aerospace Instrumentation and works as the assistant research worker at the Pavlov Institute of Physiology of RAS. Research interest is computer vision. E-mail: *tamimio.yvo@hotmail.com*

Vadim Rostislavovich Lutsiv, (b. 1954) graduated from Leningrad Institute of Aerospace Instrumentation in 1977, majoring in Electronic Computers. He received a Doctor of Technical Science degree by the decision of the dissertation council at the Saint Petersburg State University of Aerospace Instrumentation in 2012. E-mail: *vluciv@mail.ru*

Received November 3, 2023. The final version – March 13, 2024.