Распознавание визуальных частиц речи для обучения правильной артикуляции

Мурыгин К.В.,
кандидат технических наук

В статье приводятся результаты исследований по проблеме распознавания визуальных частиц речи. Целью проведения исследований является разработка программной системы обучения правильной артикуляции при произнесении речи для упрощения её понимания слабослышащими людьми. Кроме этого, результаты исследований могут использоваться для дополнения звукового информационного канала визуальным, что необходимо для повышения качества распознавания речи в условиях шума или посторонних источников звука.

Results of researches devoted to recognition of visual speech particles are described in the article. The purpose of carrying out the researches is the development of program system for training of correct articulation during speech pronouncing for simplification of its understanding by deaf and hard of hearing people. Besides, the results of researches can be used for supplementing the sound information channel with the visual one, that is necessary for improvement of speech recognition quality in the conditions of noise or extraneous sound sources.

Введение

Большинство исследований в области распознавания речи ведутся в направлении интеллектуального анализа звуковой информации, которая считается наиболее информативной при передаче речевого кода. Этот факт проявляется, в частности, в значительных затруднениях, которые испытывает при общении человек с нарушениями слуха. Кроме того, фонетический состав языка является более полным, чем его визуальный алфавит. Так, в украинской речи встречаются 6 гласных и 32 согласные

фонемы, которые визуально можно представить не более чем 16 артикуляционными образами. Кроме восприятия звуковой речи, можно отметить существование возможности чтения с губ специально подготовленными людьми. Методики обучения этим навыкам, конечно, ориентированы, в основном, на слабослышащих и глухих и предполагают хорошее владение языком, на котором происходит разговор, его структурой, знание контекста, позволяющее получать дополнительную информацию на основе смыслового комбинирования. Автоматизация процесса зрительного восприятия речи влечёт необходимость применения достаточно качественного и сложного семантического анализатора, что предполагает использование не только математического аппарата распознавания зрительных образов.

Тем не менее, приведённые сложности решения общей задачи не исключают возможности получения важных, с практической точки зрения, результатов уже на начальных этапах, связанных с обнаружением и распознаванием области губ и её конфигурации. Таким результатом может быть система обучения правильной артикуляции для облегчения зрительного восприятия устной речи людьми с нарушением слуха. В настоящее время системы автоматического чтения с губ в большинстве своём разрабатываются для дополнения звукового информационного канала визуальным. В этой связи создание обучающей системы является новым направлением и имеет значительную практическую ценность.

Задача автоматического чтения по губам объединяет в себе несколько подзадач, некоторые из которых имеют самостоятельное практическое значение, а именно:

— обнаружение лиц (face detection);

— обнаружение или извлечение деталей лица, в частности области губ (face features extraction, lip tracking);

— выделение признаков для описания конфигурации губ, позволяющих находить соответствие с произнесённой фонемой, и разработка методов такой классификации (lip reading).

По аналогии с элементарными звуковыми частицами речи – фонемами – для обозначения элементарных визуальных образов речи будем использовать уже вполне устоявшийся термин – *визема*. Внимание данной статьи будет сконцентрировано на решении задачи классификации визем.

1. Формирование словаря визем

Согласно фонетике, в украинском языке существует 6 гласных и 32 согласные фонемы:

[і], [и], [е], [у], [о], [а];

[б], [п], [д], [д'], [т], [т'], [г], [к], [ф], [ж], [з], [з'], [ш], [с], [с'], [г], [х], [дж], [дз], [дз'], [ч], [ц], [ц'], [в], [й], [м], [н], [н'], [л], [л'], [р], [р'].

Здесь знак ' означает мягкость.

Сопоставляя фонетический состав украинского языка с исследованиями Бельтюкова В.И. для русского языка [1, 2] и учитывая фонетическое сходство украинского и русского языков, можно сформировать следующий визуальный алфавит украинских звуков (визем), представленный в табл. 1.

Таблица 1

Визуальный алфавит украинских звуков (визем), полученный по аналогии с алфавитом В.И. Бельтюкова

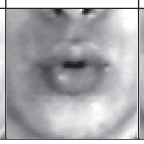
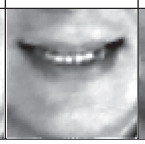
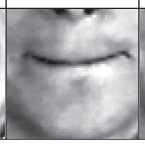
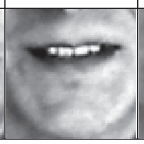
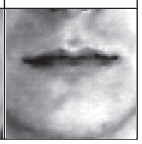
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
а	о	у	е	і	п	ф	ш	л'	р	с'	т	т'	к	й
				и	б	в	ж	р'		с	д	д'	г	
					м		ч			з'	н	н'	х	
							дж			з	л		ґ	
										ц				
										ц'				
										дз				
										дз'				

Предварительный анализ возможности автоматической классификации образов такого алфавита показал необходимость его существенного сокращения в направлении использования базовых, или опорных, визем [3]. В приведённой табл. 1 виземы, начиная с девятой, являются плохо различимыми даже человеком с его значительно более мощным зрительным аппаратом. Это во многом связано с тем, что процесс воспроизведения соответствующих им звуков в значительной мере скрыт внутри ротовой полости, что существенно усложняет их зрительное восприятие и, тем более, автоматическое распознавание на основе полученного цифрового изображения.

Поэтому в качестве рабочего алфавита визем принят алфавит, включающий в себя только опорные виземы (см. табл. 2).

Таблица 2

Принятый рабочий алфавит визем

1	2	3	4	5	6	7	8	9
а	о	у	е	і	п	ф	ш	$
				и	б	в	ж	
					м		ч	
							дж	
								

Для сравнения с более широким алфавитом (табл. 1) в таблице 3 приведены также изображения визем для элементов алфавита, не вошедших в рабочий алфавит (см. табл. 2).

Таблица 3

Элементы расширенного алфавита, не вошедшие в рабочий алфавит

9	10	11	12	13	14	15
л'	р	с'	т	т'	к	й
р'		с	д	д'	г	
		з'	н	н'	х	
		з	л		г	
		ц				
		ц'				
		дз				
		дз'				
						

Как видно из табл. 3, приведённые в ней элементы алфавита визуально трудно отличимы от элементов рабочего алфавита, что может существенно затруднить распознавание произнесённого звука по изображению соответствующей конфигурации губ. Так, виземы 10 и 12 визуально трудно отличимы от 8, а виземы 9, 11, 13 и 15 легко спутать как между собой, так и с виземой 5 принятого рабочего алфавита.

Знак $ в рабочем алфавите визем означает нормальное положение, молчание, паузу или любую другую визему, не входящую в этот алфавит. Таким образом, при распознавании предпочтение отдаётся виземам 1–8. В случае отказа от распознавания (не распознана ни одна из восьми визем) данной конфигурации приписывается значение 9, которое также может генерироваться в промежуточном положении между двумя и более виземами.

2. Используемая база данных

Несмотря на то, что для создания описанной системы обучения необходимо решить несколько задач интеллектуальной обработки визуальной информации [4], будем считать, что положение области губ предварительно определено, например, методами [5, 6, 7]. Таким образом, на вход классификатора поступают вырезанные изображения области губ.

Для наполнения базы данных изображений визем была разработана специальная программа, позволяющая вырезать изображения визем из видеопотока. Пользователь сначала выбирает визему, которую он будет вводить, при этом ему демонстрируется пример правильного произнесения (см. рис. 1). После этого он самостоятельно воспроизводит выбранную визему, стараясь добиться максимального соответствия эталону, и сохраняет результирующее изображение на диск.

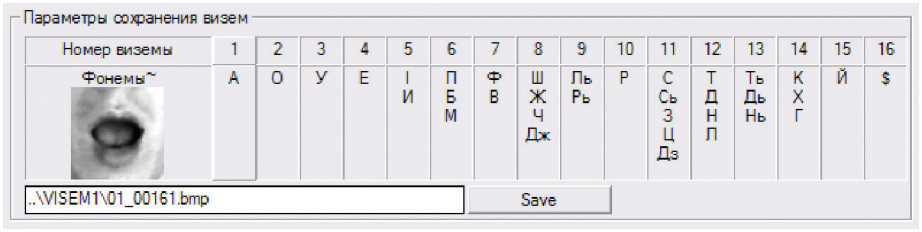


Рис. 1. Заполнение базы данных визем

С применением описанной программы была сформирована база изображений визем, произносимых разными людьми на разном расстоянии от камеры, насчитывающая 988 изображений разных размеров (см. рис. 2).

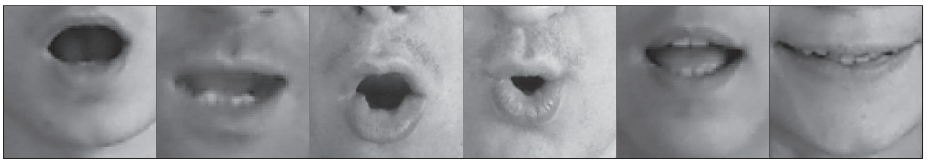


Рис. 2. Пример изображений визем из базы данных

Для исследований возможностей автоматического распознавания визем полученная база данных была размечена вручную. В ходе разметки на каждом изображении отмечались четыре точки, характеризующие крайние левое, правое, нижнее и верхнее положения точек губ. Для приведения изображений к одному масштабу за масштабный коэффициент был выбран горизонтальный размер области губ. Экспериментальная зависимость частот распределения отношения вертикального размера области губ к горизонтальному показана на гистограмме ниже (см. рис. 3).

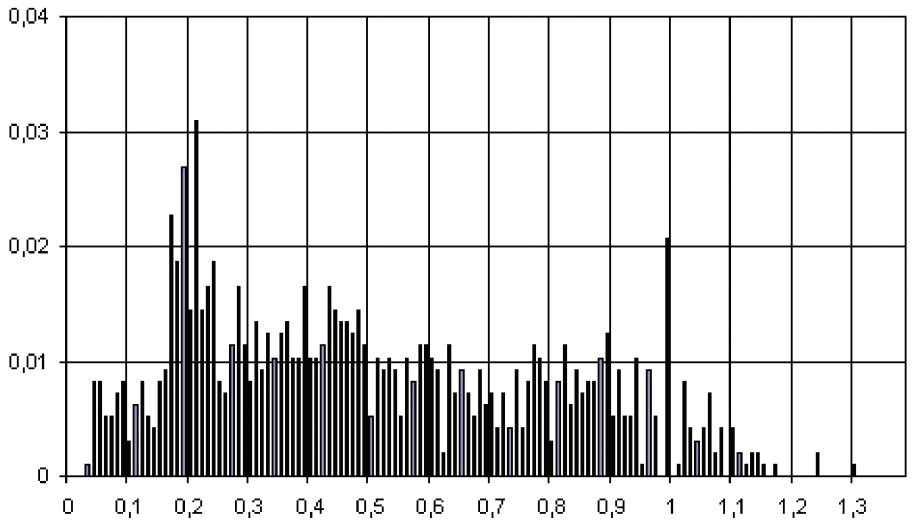


Рис. 3. Гистограмма частот отношения вертикального размера области губ к горизонтальному

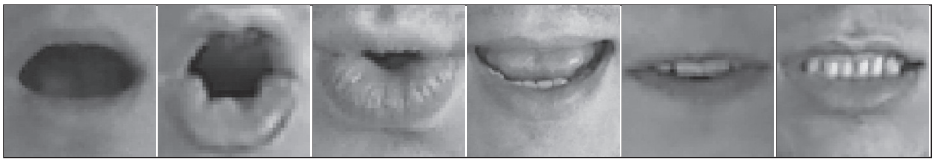


Рис. 4. Пример изображений из базы данных области губ после уточнения их положения по имеющейся разметке

Из приведённой гистограммы следует, что основная часть изображений губ сосредоточена в интервале [0;1] для отношения вертикального размера области губ к горизонтальному. Это позволяет перейти к описанию области губ в виде квадрата, ширина и высота которого равны горизонтальному размеру губ, известному для каждого изображения из базы на основе сделанной разметки, с центром в точке, задаваемой выражениями:

$$X_{ц} = \frac{X_{max} + X_{min}}{2}, \; Y_{ц} = \frac{Y_{max} + Y_{min}}{2},$$

где $X_{max}, X_{min}, Y_{max}, Y_{min}$ – соответственно крайние правые, левые, нижние и верхние координаты вручную размеченной области губ.

После обработки с учётом данных разметки на основе описанной методики получена база изображений, более точно описывающих область губ (см. рис. 4).

Полученная база допускает масштабирование с использованием стандартных методов изменения размеров изображений и удобна для использования при разработке и исследовании как методов обнаружения области губ, так и методов распознавания визем.

3. Используемые для классификации признаки

С точки зрения необходимости использования достаточно простых алгоритмов получения признаков, наиболее приемлемым является использование Хаар-подобных свойств, представляющих собой результат сравнения яркостей в двух прямоугольных областях изображения (см. рис. 5).

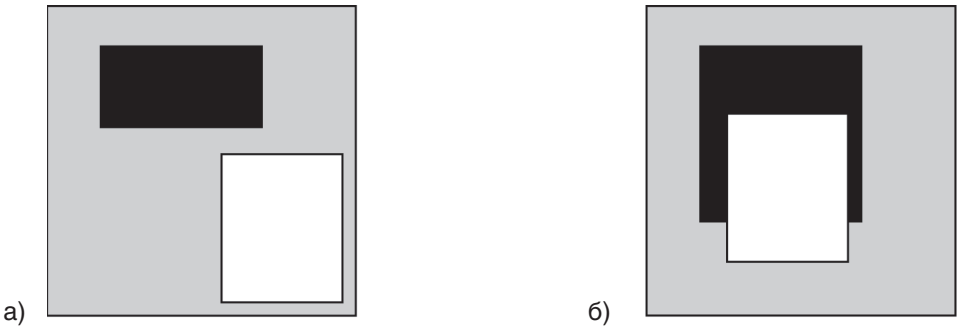


Рис. 5. Вид прямоугольных свойств, используемых в качестве признаков: а) области не пересекаются; б) области пересекаются

Значение признака для данной области изображения или отклик области изображения на данное свойство вычисляется на основе следующего выражения:

в случае непересекающихся областей –

$$R = \begin{cases} 1, \dfrac{S_Б}{N_Б} > \dfrac{S_Ч}{N_Ч}; \\ -1, \dfrac{S_Б}{N_Б} \le \dfrac{S_Ч}{N_Ч}. \end{cases},$$

в случае пересечения областей –

$$R = \begin{cases} 1, \dfrac{S_Б}{N_Б} > \dfrac{S_Ч - S_{Ч \cap Б}}{N_Ч - N_{Ч \cap Б}}; \\ -1, \dfrac{S_Б}{N_Б} \le \dfrac{S_Ч - S_{Ч \cap Б}}{N_Ч - N_{Ч \cap Б}}. \end{cases}$$

Здесь индексы Ч и Б обозначают чёрную и белую области соответственно, а Ч∩Б – область пересечения областей чёрного и белого цвета; S – сумма яркостей пикселей изображения, находящихся под областью; N – число пикселей изображения, находящихся под областью. Значения, получаемые на основе этих выражений, являются инвариантными по отношению к любым линейным преобразованиям функции яркости изображений, к которым с достаточной точностью можно отнести операции изменения яркости и контраста.

4. Обучение и тестирование классификаторов визем

Для решения задачи классификации конфигурации губ на входных изображениях, согласно введённому рабочему алфавиту визем, использовался подход, основанный на группировке описанных выше признаков в классификаторы с использованием алгоритма AdaBoost. Переход от задачи разделения двух классов был решён путём построения набора классификаторов, отделяющих каждую из визем принятого алфавита (см. табл. 2) от всех остальных визем. Такой подход позволяет получить больше информации о распознаваемом объекте за счёт возможности множественной классификации. При множественной классификации объект, поступивший на вход распознавателя, относится сразу к нескольким классам визем. Это позволяет контролировать и использовать промежуточные, переходные состояния между виземами, что может быть очень полезно при решении задачи автоматического анализа последовательности визем – слов или предложений в процессе слитной речи.

Каждое изображение базы данных визем было дополнено зеркально отражённым в горизонтальном направлении изображением. Для уменьшения влияния масштаба изображения на входе на значения откликов

прямоугольных свойств (см. рис. 5) изображения в процессе обучения дополнялись набором тех же изображений различного масштаба. Для экспериментов использовались 30 масштабов обучающих изображений в диапазоне 30–90 пикселей с шагом 2 пикселя. Используемый диапазон масштабов является характерным для входных изображений области губ, получаемых на выходе алгоритма поиска губ. Для объективного тестирования полученных результатов обучения используемая база изображений визем была разделена на две части – обучающий и тестовый наборы. На обучающем наборе проводилось обучение классификаторов. На тестовом наборе, не пересекающемся с обучающим, проводилось тестирование полученных классификаторов.

Процесс обучения по методу AdaBoost проводился до достижения классификатором средней ошибки классификации менее 0.001. Количество используемых прямоугольных признаков, необходимых для достижения указанной ошибки, в экспериментах не превосходило 50. С учётом того, что входными данными для распознавания является область изображения, заключающая в себе предварительно обнаруженные губы, можно сделать вывод о высокой скорости обработки данных и об отсутствии необходимости её увеличения за счёт использования каскада классификаторов.

Тестирование полученных классификаторов на тестовом наборе показало результаты, приведённые в табл. 4.

Таблица 4

Матрица принятых решений на тестовом наборе

Входные виземы	Распознанные виземы									Множественное распознавание или отказ от распознавания
	1	2	3	4	5	6	7	8	9	
	а	о	у	е	і	п	ф	ш	$	
					и	б	в	ж		
						м		ч		
								дж		
										
	66	-	-	2	-	-	-	-	-	12
	-	76	-	-	-	-	-	-	-	8
	-	-	103	-	-	-	-	-	1	8

Входные виземы	Распознанные виземы									Множественное распознавание или отказ от распознавания
	1	2	3	4	5	6	7	8	9	
	а	о	у	е	і	п	ф	ш	$	
	-	-	-	28	1	-	-	2	-	3
	-	-	-	-	14	-	-	-	-	-
	-	-	-	-	-	38	-	-	2	6
	-	-	1	-	-	-	41	-	-	2
	-	-	-	-	-	-	-	33	-	9
	-	-	-	-	-	-	-	-	93	7

Приведённые в табл. 4 данные говорят о достижении удовлетворительных результатов по распознаванию визем. Наибольшие ошибки связаны с множественным распознаванием и отказом от распознавания, что объясняется влиянием индивидуальных визуальных особенностей артикуляционного аппарата различных людей. В приведённой матрице принятых решений, полученной на тестовом наборе, не выявлено устойчивой неправильной классификации каких-либо двух классов, что свидетельствует об отсутствии необходимости сокращения принятого алфавита классов-визем путём объединения плохо разделимых классов в один. Достигнутые показатели качества позволили использовать описанный подход при разработке программы обучения правильной артикуляции для облегчения восприятия устной речи людьми с нарушениями слуха.

Литература

1. *Нейман Л.В.* Анатомия, физиология и патология органов слуха и речи. М.: Просвещение, 1977 г. 146 с.

2. Методика обучения глухих устной речи: Учеб. пособие для студентов дефектол. фак. фед. ин-тов. / Под ред. проф. Ф.Ф. Рау. М.: Просвещение, 1976. 279 с.

3. http://www.pedlib.hut.ru/Books/pravdina/pravdinap124.html.

4. *Мурыгин К.В.* Концепция системы распознавания речи на основе чтения по губам // Искусственный интеллект. 2009. №2. С. 116–123.

5. *Paul Viola and Michael J. Jones*. Robust real-time object detection //In Proc. of IEEE Workshop on Statistical and Computational Theories of Vision, 2001.
6. *Мурыгин К.В.* Обнаружение объектов на изображении на основе каскада классификаторов // Искусственный интеллект. 2007. №2. С. 104–108.
7. *Мурыгин К.В.* Особенности реализации алгоритма AdaBoost для обнаружения объектов на изображениях // Искусственный интеллект. 2009. №3. С. 573–581.

Мурыгин К.В.
Кандидат технических наук, начальник отдела распознавания зрительных образов Института проблем искусственного интеллекта (г. Донецк, Украина).
Область научных интересов: разработка методов и алгоритмов интеллектуального анализа зрительной информации. Занимался вопросами поиска объектов на изображении, распознавания человека по изображению лица, анализа движения в видеопоследовательности, стереозрения – восстановления карты диспаратности по набору изображений.
info@iai.dn.ua