Электронный журнал «Техническая акустика»
http://www.ejta.org

2014, 9

В. В. Митянок

Полесский государственный университет, Беларусь 225710, Пинск, ул. Днепровской флотилии, 23, e-mail: mitsianok@mail.ru

О физической структуре звуков З, ЗЬ, Ж, ЖЬ

Получена 26.09.2014, опубликована 04.11.2014

Метод аппроксимации применяется для разложения звуков З, ЗЬ, Ж, ЖЬ на моды с различными несущими частотами. Показано, что среди мод есть ведущие себя регулярно, есть хаотически вспыхивающие и тут же гаснущие, есть модулированные частотой первой из регулярных мод. Высказано предположение о том, что такое поведение мод приводит к явлению формант. Дано объяснение трудностям методов распознавания речи, основанных на преобразованиях Фурье. Предложено несколько вероятностных критериев, как для различения вышеуказанных звуков, так и для различения людей, произносящих эти звуки.

Ключевые слова: автоматическое распознавание речи, цифровая обработка сигналов, распознавание человеческого голоса.

ВВЕДЕНИЕ

При решении задачи автоматического распознавания речи человека должны быть пройдены следующие этапы. Должен быть разработан аппарат распознавания отдельных звуков (фонем), слогов, групп слогов, образующих слова, предложений, составленных из этих слов, осмысление слов и предложений, перевод услышанного в грамматически выверенный текст. Эти проблемы не обязательно должны разрабатываться последовательно, вполне допустим и параллельный ход, однако недочеты в решении хотя бы в одной из перечисленных проблем пагубно скажутся на решении задачи автоматического распознавания речи в целом. Надо сказать, что в настоящее время ни одна из вышеперечисленных проблем еще не решена «под ключ», хотя и имеются несомненные достижения.

Имеющиеся достижения и еще не решенные проблемы весьма полно изложены в [1-3]. В [4] можно найти обширный перечень направлений, тем и сопутствующих вопросов, имеющих отношение к задаче автоматического распознавания речи человека. В [5] даже утверждается, что «распознавание речи на акустическом и фонетическом уровне в настоящее время доведено до совершенства, то есть сравнимо по качеству с надежностью распознавания отдельных звуков человеком», что, скорее всего, является некоторым преувеличением. Авторы [5], надо полагать, имели в виду звуки, произносимые в неких идеальных условиях: четкость дикции, полное отсутствие шумов и т. д. Напротив, в [3] отмечено, что «практика применения систем

распознавания речи показала, что они неустойчивы к помехам и искажениям канала речевой связи. Типичным является катастрофическое снижение словесной надежности распознавания до величин порядка 40-60% при появлении относительно слабых шумов, смене типа микрофона или канала связи». Свидетельством тому, что этот узел проблем еще далек от завершения, является также тот факт, что в почти каждом выпуске издающегося с 2008 года журнала «Речевые технологии» публикуется 1-2 а то и более статей, относящихся к фонемам, формантам, характеристикам отдельных звуков. Не говоря уже о других изданиях.

С задачей автоматического распознавания речи тесно связаны задача распознавания личности человека по голосу и задача компьютерного синтеза и клонирования речи. Имеющиеся достижения и проблемы компьютерного синтеза и клонирования речи, включая историю вопроса, можно найти в [6]. Там, в частности, описана система фонем русского языка, аппарат речеобразования человека и ряд сопутствующих вопросов, изложены основы теории клонирования персональных характеристик речи. В [7] с той же целью излагается сравнение фонетических систем белорусского и русского языков.

На первых порах большие трудности представлялись в распознавании *слитной* речи, но ряд удачных решений позволил и здесь достигнуть определенного прогресса. В частности, представляет интерес подход, основанный на различного рода вероятностных моделях прогноза распознавания акустических образов [5].

Среди звуков, произносимых человеком, более простыми являются те, которые могут произноситься долго – столько, на сколько хватает дыхания. В свою очередь, эти звуки можно разделить на низкочастотные, например, звуки А, О, Э, У, И, Ы и высокочастотные: З, ЗЬ, Ж, ЖЬ, С, СЬ, Ш, ШЬ и т. д. Первые 6 низкочастотных звуков из числа перечисленных выше могут быть исследованы различными методами, в том числе и методом преобразований Фурье. В [8] методом аппроксимации [9, 10] было проведено исследование фазовых комбинаций различных мод этих звуков и было обнаружено, что существуют такие комбинации, которые являются уникальными как для отдельных респондентов, так и для отдельных звуков, что может быть использовано для создания систем распознавания личности по голосу. В связи с этим возникла идея найти подобные фазовые комбинации также и в отношении высокочастотных звуков. К сожалению, эти надежды не оправдались, тем не менее, получен ряд результатов, которые могут быть полезными для решения проблем распознавания речи и идентификации диктора. Они излагаются ниже.

1. ИСХОДНЫЕ ПОЛОЖЕНИЯ

В настоящей статье приведены результаты исследования звуков З, ЗЬ, Ж, ЖЬ, поскольку между ними есть много общего. Для исследования использован метод аппроксимации [9, 10]. Согласно этому методу некая функция $y(t)$, заданная своими значениями в n точках $t_i : i=1..n$ и про которую также известно, что она представляет собой сумму (почти) гармонических функций, может быть разложена на исходные

слагаемые (моды) путем составления невязки определенного вида и минимизации ее. В результате получается разложение

$$y_i = y(t_i) = b_{0i} + \sum_{j=1}^{l} a_{ji} \sin(\omega_j i) + \sum_{j=1}^{j} b_{ji} \cos(\omega_j i)\,,\ i=1..n, \qquad (1)$$

где b_{0i} – дрейфующее начало отсчета (дрейфующий нуль), a_{ji}, b_{ji} – дрейфующие синус- и косинус-амплитуды различных мод, l – количество мод, ω_i – их частоты, так называемая ловящая сеть [8]. В (1) и всюду ниже для простоты принято $t_i = i$.

Звуки, полученные от 9 респондентов, вводились в компьютер через бытовой микрофон при частоте дискретизации 44100 Гц. Если каждому новому отсчету дискретизации соответствует увеличение i в (1) на 1, то значение ω=1 соответствует физической частоте $44100/(2\pi)$ Гц≈7019 Гц. Всюду ниже именно это значение используется в качестве единицы частоты. Периоды будем измерять в обратных единицах: $T = (2\pi)/\omega$.

От (1) нетрудно перейти к физически более информативной записи

$$y_i = y(t_i) = b_{0i} + \sum_{j=1}^{l} c_{ji} \sin(\omega_j i + \varphi_{ji})\,,\ i=1..n, \qquad (2)$$

где c_{ji} – *общие* амплитуды мод, φ_{ji} – их фазы. Амплитуды и фазы связаны между собой очевидными соотношениями

$$c_{ji} = \sqrt{a_{ji}^2 + b_{ji}^2}\,, \qquad \varphi_{ji} = arctg\left(\frac{b_{ji}}{a_{ji}}\right), \qquad i=1..n,\ \ j=1..l, \qquad (3)$$

$$a_{ji} = c_{ji}\cos(\varphi_{ji})\,, \qquad b_{ji} = c_{ji}\sin(\varphi_{ji})\,, \qquad i=1..n,\ \ j=1..l. \qquad (4)$$

Прежде всего, для всех полученных образцов вычислялся спектр методом преобразований Фурье. Во всех случаях один из максимумов спектрограммы был виден особенно четко – для разных респондентов и разных звуков его частота находилась в интервале 0.016-0.032 (в наших единицах). Эту частоту естественно назвать базовой, а соответствующий ей период – базовым. Ловящая сеть выбиралась следующим образом. Первая группа частот – это базовая частота и 3 кратных ей обертона. Далее – частоты 0.188, 0.215, 0.35 – вторая группа частот, они также часто просматривались на спектрограммах, и, затем, эквидистантный набор частот от 0.512 до 2.564 с шагом 0.108 – третья группа. Использовались также и другие варианты выбора частот 3-й группы, но принципиальной разницы с излагаемыми ниже результатами обнаружено не было. Отметим также в этой связи, что точный выбор частот ловящей сети не очень важен для решения задачи аппроксимации, и это является одним из достоинств метода. В самом деле, любая из мод, входящих в (1) и имеющая несущую частоту ω_j, легко может быть перелицована на моду с близкой частотой ω_j': $\omega_j = \omega_j' + \Delta\omega$ путем тривиальных преобразований

$$a_{ji}\sin(\omega_j i)+b_{ji}\cos(\omega_j i)=a_{ji}\sin(\omega_j^{'}i+\Delta\omega i)+b_{ji}\cos(\omega_j^{'}i+\Delta\omega i)=$$
$$=a_{ji}^{'}\sin(\omega_j^{'}i)+b_{ji}^{'}\cos(\omega_j^{'}i), \qquad (5)$$

где

$$a_{ji}^{'}=a_{ji}\cos(\Delta\omega i)-b_{ji}\sin(\Delta\omega i),\quad b_{ji}^{'}=a_{ji}\sin(\Delta\omega i)+b_{ji}\cos(\Delta\omega i),\quad i=1..n,\ j=1..l. \qquad (6)$$

Нетрудно показать, что перелицованные общие амплитуды, вычисляемые по первой из формул (3), от такого преобразования вообще не меняются, меняются только фазы.

2. ОСОБЕННОСТИ МОД

В результате первичного разложения звуковых функций на моды оказалось, что выбор частот первой группы нуждается в корректировке. Так, было обнаружено, что на графики общих амплитуд, синус- и косинус- амплитуд первой моды ловящей сети накладываются некие новые колебания с частотой, равной половине базовой. Равным образом во многих случаях на дрейфующие амплитуды моды №2 накладываются новые колебания с частотой, равной 1.5 базовой. Следовательно, в сеть следует включить также новые частоты, равные 0.5, 1.5, 2.5, 3.5 базовой. Это означает, что та частота, которая ранее была первой, теперь является частотой №2, та частота, которая была частотой №2, теперь является частотой №4 и т. д. Тем не менее, сохраним за частотой №2 *исправленной* сети термин «базовая», поскольку именно соответствующая ей мода показывает наиболее регулярное поведение. Новые, добавленные частоты будем называть полуцелыми от базовой. На рис. 1 изображено поведение общих амплитуд (красные линии), синус- и косинус- амплитуд (соответственно синие и зеленые линии) моды, соответствующей базовой частоте и рассчитанных ловящей сетью, не учитывающей полуцелые частоты (верхняя группа кривых), и сетью, учитывающей полуцелые частоты (нижняя группа кривых).

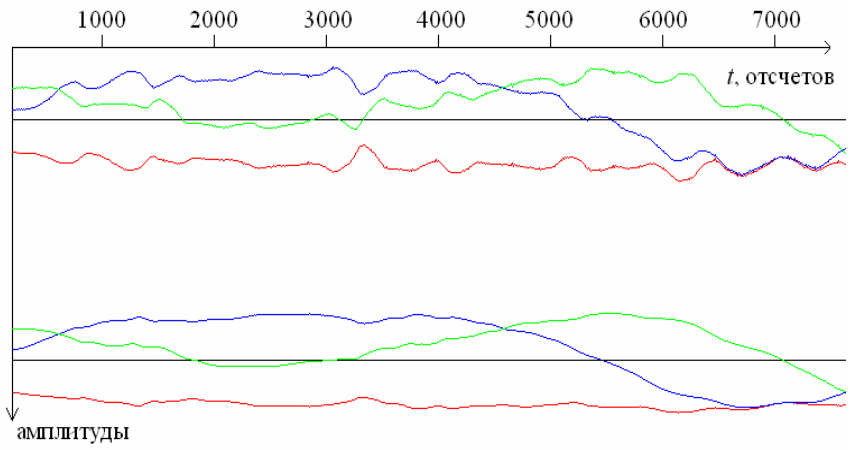

Рис. 1. Поведение общих амплитуд (красные линии), синус- и косинус- амплитуд (соответственно синие и зеленые линии) моды, соответствующей базовой частоте для ловящей сети, не учитывающей полуцелые частоты (верхняя группа кривых), и сети, учитывающей полуцелые частоты (нижняя группа кривых)

Для удобства рассмотрения, группы амплитуд разнесены по вертикали. Черная горизонтальная прямая посередине каждой из групп – ее начало отсчета. Время по горизонтальной линии оси координат измеряется в отсчетах дискретизации, они составляют 1/44100 долю секунды.

Графики рисунка 1 получены для звука ЗЬ от респондента №9. Для других звуков и (или) респондентов ситуация выглядит аналогично. Сравнивая поведение двух групп амплитуд на рисунке 1 нетрудно заметить, что нижняя группа кривых является более «аккуратной», что является свидетельством реальности полуцелых частот. Полуцелые частоты не распознаются на спектрограмме Фурье, но метод аппроксимации позволяет установить их существование. Ниже, на рисунке 2, представлены 5 первых мод звука З в исполнении респондента №1, полученных сетью, учитывающей полуцелые частоты.

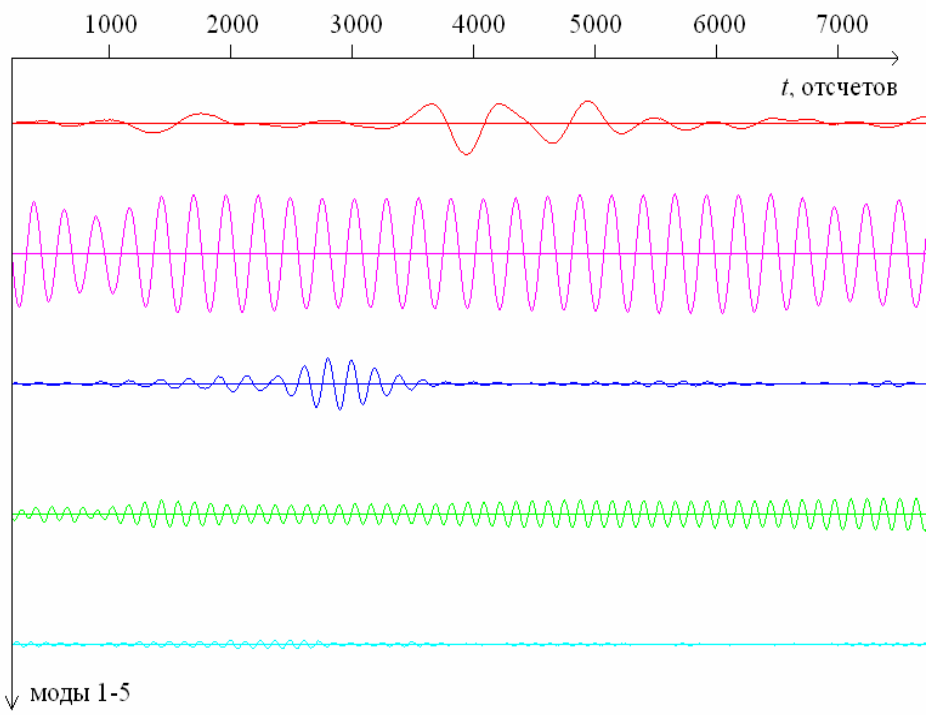

Рис. 2. Моды 1-5 (сверху вниз, по порядку) звука З в исполнении респондента №1. Для удобства рассмотрения моды разнесены по вертикали. Горизонтальная прямая того же цвета посередине каждой из мод – ее начало отсчета. Время по горизонтальной линии оси координат измеряется в отсчетах дискретизации

Как видно из рисунка 2, моды №2 и №4 ведут себя регулярно, моды №1, №3, №5, соответствующие полуцелым частотам, проявляются как бы «вспышками». В появлении вспышек не прослеживается никакой системы, похоже на то, что они случайны. В максимуме их интенсивность может быть значительной, однако из-за кратковременности они не могут сформировать четкие линии на спектрограмме Фурье.

Отметим, что при обратном суммировании мод в звуковую функцию дрейфующий нуль и моды, соответствующие полуцелым частотам, могут быть опущены. На слух звук остается тем же.

Оказалось также, что существуют такие частоты, для которых соответствующие им моды малоинтенсивны в *любой* момент времени. Поэтому эти частоты могут быть опущены из ловящей сети. Уменьшение количества мод ловящей сети приводит к значительному сокращению времени расчетов компьютера. Опускаемые частоты одинаковы для всех 9 респондентов, что позволяет выдвинуть предположение, что они одинаковы для всех людей вообще.

Далее, наблюдался эффект синхронизации общих амплитуд высших мод модой №2. (рис. 3)

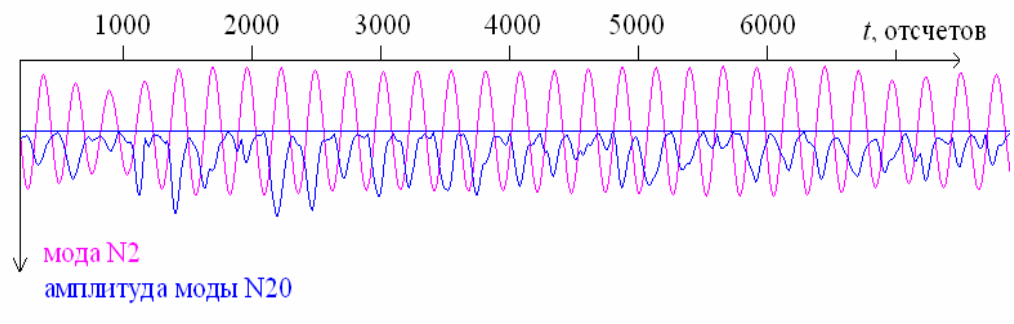

Рис. 3. По горизонтальной оси – номер точки дискретизации. По вертикальной оси: красная линия – мода №2, синяя линия – *общая* амплитуда моды №20 (ω=1.376). Произносится звук ЗЬ в исполнении респондента №2. Для удобства рассмотрения вертикальный масштаб общей амплитуды моды №20 сжат в 3 раза

Из рисунка 3 видно, что мода №2 как бы задает ритм общей амплитуде моды №20, синхронизирует ее. На каждом пульсе (участке между двумя соседними максимумами, или, как вариант, минимумами) моды №2 общая амплитуда моды №20 то достигает почти нулевого значения, то достигает максимального значения, фактически имеет место ее модуляция частотой, равной базовой частоте. При этом форма пульсов общей амплитуды моды №20 заметно меняется при продвижении по оси времени. Имеют место как бы срывы (неустойчивость) пульсаций, однако после срывов опять вступает в силу ритм, задаваемый модой №2, и пульсации восстанавливаются. До следующего срыва. Модуляция со срывами наблюдается для всех изученных звуков, всех респондентов и всех мод с частотами, большими, чем 0.5. Из рисунка 3 становятся понятными причины трудностей методов распознавания речи, основанных на преобразованиях Фурье: из-за модуляции, срывов и неустойчивости, сравнимых по величине с параметрами мод, спектр становится очень сложным для расшифровки. Надо полагать, что именно эти причины приводят к появлению такого понятия, как форманты.

3. СТАТИСТИКА БАЗОВЫХ ПЕРИОДОВ

Мода №2 показывает регулярное поведение для всех респондентов и всех звуков. Однако ее периоды (базовые периоды) могут заметно отличаться от респондента к респонденту и от звука к звуку. Приведем таблицу доверительных интервалов для базовых периодов. При проведении данного исследования сеансы каждого звучания разрезались на фрагменты длиной около 3000 точек, измерялся средний базовый период по каждому из фрагментов. Всего таким способом от каждого из респондентов было получено от 136 до 410 чисел − значений базовых периодов каждого из звуков.

Табл. 1. Доверительные интервалы значений базового периода

Респондент	Звук			
	З	ЗЬ	Ж	ЖЬ
1	276.2-289.5	324.1-330.2	304.4-310.6	330.2-339.6
2	320.6-332.9	313.0-323.2	365.4-375.8	288.1-316.6
3	220.2-227.3	210.5-222.9	237.0-242.3	232.3-239.4
4	284.9-292.5	260.2-267.7	288.7-301.5	249.5-257.6
5	369.4-375.9	359.3-367.5	356.8-364.9	350.2-360.0
6	365.7-369.1	375.7-378.2	360.6-363-8	374.1-377.0
7	285.7-290.0	316.7-321.1	329.1-333.4	326.2-331.7
8	203.5-206.0	218.1-220.8	211.8-215.5	210.6-212.1
9	222.7-228.6	226.0-229.3	224.8-228.5	197.0-199.6

В пределах одного сеанса звучания (несколько секунд) базовый период может немного меняться. Можно найти среднее по сеансу значение периода, среднее квадратичное отклонение. Для экономии места здесь приведены лишь доверительные интервалы. Во всех случаях величина усредненного по фрагменту базового периода попадает в указанные в таблице 1 интервалы с вероятностью около 50 процентов. Если же для каждого из респондентов и каждого из звуков произвести отброс 20 процентов результатов, наиболее отличающихся от среднего значения, то вероятность попадания в указанный интервал составляет уже от 50 до 75 процентов – в среднем 63 процента.

Как видно из таблицы 1, если известен произносимый звук, то по измеренному базовому периоду можно сделать определенные суждения относительно того, кем этот звук был произнесен, и, наоборот, если известен респондент, то можно иметь определенные суждения о том, какой именно из звуков был произнесен. Но не всегда. Из таблицы 1 видно, что, например, для звука ЗЬ перекрываются интервалы для респондентов №3 и №8, а также респондентов №2 и №7, близки между собой интервалы респондентов №1 и №2, интервалы же для остальных респондентов находятся достаточно далеко друг от друга. Также видно, что для звука З перекрываются зоны респондентов №3 и №9, респондентов №1, №4, №7 и респондентов №5 и №6. Для звука Ж – перекрываются зоны респондентов №5 и №6, для звука ЖЬ – частично перекрываются зоны респондентов №1 и №7.

Тем не менее, из таблицы 1 всегда можно извлечь определенную пользу для решения задачи распознавания. Например, для респондентов №1 и №2, доверительные интервалы которых для звука ЗЬ близки, можно выдвинуть критерий – если наблюдаемый период превосходит значение 323.65 (середина зазора между интервалами) то звук ЗЬ был произнесен респондентом №1, в противном случае – респондентом №2. Приведем таблицу результатов различения респондентов №1 и №2 по данному критерию.

Табл.2. Надежность распознавания респондентов №1 и №2, произносящих звук ЗЬ

Звук произносит	Показано на респондента		Надежность распознавания	Всего образцов
	Респондент №1	Респондент №2		
Респондент №1	227 случаев	62 случая	78.5%	289
Респондент №2	115 случаев	295 случаев	72%	410

Если же отбросить 20% образцов, для которых базовые периоды наиболее отличаются от среднего, то результаты таковы:

Табл.3. Надежность распознавания респондентов №1 и №2, произносящих звук ЗЬ, при отбрасывании 20% образцов, показывающих наибольшее отклонение от среднего значения

Звук произносит	Показано на респондента		Надежность распознавания	Всего образцов
	Респондент №1	Респондент №2		
Респондент №1	199 случаев	33 случая	85.8%	232
Респондент №2	64 случая	264 случая	80.5%	328

Таким образом, даже в случае близости доверительных интервалов для респондентов №1 и №2 можно говорить об их различении с какой-то вероятностью. Более надежным для различения респондентов №1 и №2 выглядит использование вместо звука ЗЬ других звуков: З, Ж и ЖЬ. Все это справедливо и для других респондентов. По *совокупности* базовых периодов 4-х звуков можно отличить любого из 9 респондентов от всех остальных, вместе взятых, с надежностью более 85 процентов.

Рассмотрим теперь статистику значений базовых периодов. Как правило, гистограмма базовых периодов для каждого сеанса звучания представляет собой достаточно гладкую линию, близкую к кривой нормального распределения. Однако в некоторых случаях имеются заметные отклонения. Так, для звука ЖЬ и респондента №2 гистограмма имеет вид, представленный на рис. 4.

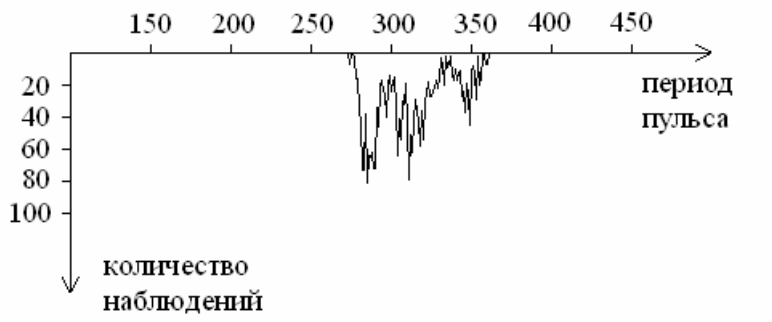

Рис. 4. Гистограмма базовых периодов для респондента №2 и звука ЖЬ. Период измеряется в отсчетах дискретизации ввода

Как видно из рисунка 4, наблюдаемые периоды разбились на 3 группы со значениями от 275 до 300, от 300 до 336 и от 336 до 360. Причем все 3 группы наблюдались даже во время одного сеанса звучания. Для других звуков и (или) других наблюдателей также иногда наблюдались похожие гистограммы. Всего для 4-х звуков и 9 респондентов, то есть 36 комбинаций наблюдалось 2 случая разбиения периодов на 3 группы, 6 случаев – разбиения на 2 группы, в остальных случаях разбиений не было. Сам факт разбиения периодов на группы также может быть использован для решения задач распознавания человека по голосу и распознавания его речи.

4. СДВИГ ПО ВРЕМЕНИ МЕЖДУ РЕГУЛЯРНЫМИ МОДАМИ

Кроме моды №2 регулярное поведение показывают также и другие четные моды, от моды №2 до моды №8. При этом во всех случаях моменты максимумов этих мод заметно отличаются. Представляет интерес вопрос о том, насколько велико такое отличие.

Отличие может быть охарактеризовано отношением задержки первого максимума изучаемого обертона (сразу после очередного максимума моды №2) к длительности этого же пульса моды №2 (относительной задержке). Как оказалось, в большинстве случаев (для разных четных обертонов, звуков и респондентов) гистограммы распределения относительных задержек близки к нормальному распределению. Для таких распределений были подсчитаны их средние значения и дисперсии и затем найдены доверительные интервалы, попадание в которые осуществляется с вероятностью, близкой к 50 процентам. Приведем их в нижеследующих таблицах 4-6.

Табл. 4. Доверительные интервалы запаздывания первых максимумов моды №4 по отношению к максимумам моды №2, выраженные в долях длительности пульса моды №2

Респондент	Звук			
	З	ЗЬ	Ж	ЖЬ
1	0.253-0.280	0.262-0.290	0.231-0.285	0.177-0.240
2	0.222-0.291	0.260-0.287	0.221-0.252	0.239-0.275
3	0.208-0.227	0.222-0.274	0.228-0.248	0.240-0.264
4	0.272-0.350	0.165-0.233	0.310-0.408	0.219-0ю287
5	0.220-0.299	0.276-0.355	0.247-0.357	0.203-0.294
6	0.227-0.238	0.245-0.270	0.175-0.208	0.217-0.236
7	0.199-0.246	0.232-0.254	0.246-0.284	0.237-0.276
8	0.234-0.263	0.248-0.267	0.235-0.259	0.278-0.307
9	0.130-0.235	0.167-0.226	0.202-0.233	0.198-0.235

Табл. 5. Доверительные интервалы запаздывания первых максимумов моды №6 по отношению к максимумам моды №2, выраженные в долях длительности пульса моды №2

Респондент	Звук			
	З	ЗЬ	Ж	ЖЬ
1	0.070-0.113	0.052-0.120	0.056-0.089	
2		0.029-0.100	0.006-0.158	0.034-0.090
3	0.074-0.120	0.095-0.235	0.026-0.228	0.027-0.054
4		0.033-0.061	0.062-0.195	0.037-0.114
5	0.063-0.098	0.096-0.161	0.103-0.142	
6	0.047-0.058	0.072-0.124	0.001-0.063	0.021-0.050
7		0.027-0.083		0.032-0.090
8		0.055-0.126	0.040-0.070	
9				0.118-0.168

Табл. 6. Доверительные интервалы запаздывания первых максимумов моды №8 по отношению к максимумам моды №2, выраженные в долях длительности пульса моды №2

Респондент	Звук			
	З	ЗЬ	Ж	ЖЬ
1	0.035-0.082		0.034-0.124	
2	0.122-0.197	0.051-0.133	0.131-0.197	0.038-0.084
3	0.018-0.065			
4	0.107-0.157	0.100-0.155	0.146-0.232	0.121-0.156
5	0.151-0.211	0.099-0.173		0.123-0.0.188
6	0.023-0.116	0.129-0.187	0.163-0.189	0.171-0.211
7	0.090-0.132	0.101-0.160	0.158-0.198	0.079-0.141
8		0.047-0.146		0.083-0.186
9	0.103-0.186	0.126-0.216		0.175-0.221

В тех случаях, когда клетки таблиц остались незаполненными, распределения заметно отличаются от нормального: они скорее могут быть представлены прямой суммой 2-х или 3-х нормальных распределений с различными математическими средними и дисперсиями (подобно рис. 4), и поэтому прямое вычисление их средних значений и дисперсий не характеризует компактности критериев в должной мере. Во всех остальных случаях реально наблюдаемое запаздывание попадает в доверительный интервал с вероятностью около 50%. Запаздывание мод также может использоваться в качестве вероятностного критерия, как различения респондентов, так и различения произнесенных ими звуков.

5. СООТНОШЕНИЕ ИНТЕНСИВНОСТЕЙ РАЗЛИЧНЫХ МОД

Разложение звукового сигнала на отдельные моды позволяет вычислить их парциальные интенсивности и по соотношению интенсивностей найти критерии различения звуков и респондентов. При этом обнаружились, что существуют критерии, которые позволяют различать разные звуки для всех респондентов *одновременно*. Вновь разобьем звуковые кривые на фрагменты длительностью около 3000 точек, выделим в каждом из них участок от первого до последнего максимума моды №2, и интенсивности всех мод усредним по выделенному участку. Критерии примем в виде $\frac{I_m}{I_n}$, где I_m, I_n – усредненные интенсивности мод с номерами *m* и *n*.

Табл. 7. Критерии интенсивности и их пороговые значения для попарного различения звуков, для всех 9 респондентов, вместе взятых. Если для некоторого образца наблюдаемое значение критерия превышает указанное пороговое, то это означает, что звучит звук, указанный в столбце данной клетки, в противном случае – в строке

	З	ЗЬ	Ж	ЖЬ
З			I_{11}/I_{29}, 13.0	I_{11}/I_{29}, 12.1
ЗЬ	I_9/I_{11}, 0.65		I_6/I_{28}, 7.8	I_{12}/I_{27}, 4.8
Ж				
ЖЬ			I_6/I_{27}, 4.6	

Для более детального ознакомления с эффективностью критериев, приведем таблицу результатов по каждому из респондентов по отдельности.

Табл. 8. Различение звуков З и ЗЬ. Столбец №2 – количество случаев когда при звучащем звуке З значение критерия превышает пороговое значение. Столбец №3 – количество случаев когда при звучащем звуке З значение критерия не достигает порогового значения. Столбец №4 – количество случаев, когда при звучащем звуке ЗЬ значение критерия превышает пороговое значение. Столбец №5 – количество случаев, когда при звучащем звуке ЗЬ значение критерия не достигает порогового значения. Использован критерий I_9/I_{11} , его пороговое значение – 0.65

Респондент	Звук З, критерий больше порогового значения	Звук З, критерий меньше порогового значения	Звук ЗЬ, критерий больше порогового значения	Звук ЗЬ, критерий меньше порогового значения
1	339	62	1	288
2	265	0	0	410
3	201	0	0	179
4	214	35	0	179
5	259	1	60	147
6	176	0	7	243
7	136	0	0	223
8	160	0	7	188
9	148	1	8	222

Табл. 9. Различение звуков Ж и ЖЬ. Использован критерий I6/I27, его пороговое значение – 4.6

Респондент	Звук Ж, критерий больше порогового значения	Звук Ж, критерий меньше порогового значения	Звук ЖЬ, критерий больше порогового значения	Звук ЖЬ, критерий меньше порогового значения
1	170	11	38	280
2	383	0	147	157
3	147	1	19	207
4	108	81	0	178
5	137	119	8	279
6	214	0	70	178
7	120	52	104	206
8	244	19	0	179
9	126	73	0	220

Табл. 10. Различение звуков З и Ж. Использован критерий I11/I29, его пороговое значение – 13.0

Респондент	Звук З, критерий больше порогового значения	Звук З, критерий меньше порогового значения	Звук Ж, критерий больше порогового значения	Звук Ж, критерий меньше порогового значения
1	0	401	175	6
2	0	265	304	79
3	59	142	148	0
4	0	249	189	0
5	0	260	256	0
6	0	176	214	0
7	0	136	172	0
8	0	160	263	0
9	0	149	199	0

Табл. 11. Различение звуков ЗЬ и ЖЬ. Использован критерий I12/I27 , его пороговое значение – 4.8

Респондент	Звук ЗЬ, критерий больше порогового значения	Звук ЗЬ, критерий меньше порогового значения	Звук ЖЬ, критерий больше порогового значения	Звук ЖЬ, критерий меньше порогового значения
1	35	254	318	0
2	13	397	230	74
3	58	121	226	0
4	0	179	178	0
5	0	207	287	0
6	0	250	248	0
7	0	223	310	0
8	0	195	179	0
9	0	230	204	16

Табл. 12. Различение звуков ЗЬ и Ж. Использован критерий I6/I28 , его пороговое значение – 7.8

Респондент	Звук ЗЬ, критерий больше порогового значения	Звук ЗЬ, критерий меньше порогового значения	Звук Ж, критерий больше порогового значения	Звук Ж, критерий меньше порогового значения
1	0	289	152	29
2	140	270	383	0
3	37	142	148	0
4	0	179	119	70
5	0	207	252	4
6	0	250	214	0
7	6	217	172	0
8	0	195	257	6
9	0	230	199	0

Табл. 13. Различение звуков З и ЖЬ. Использован критерий I11/I29 , его пороговое значение – 12.1

Респондент	Звук З, критерий больше порогового значения	Звук З, критерий меньше порогового значения	Звук ЖЬ, критерий больше порогового значения	Звук ЖЬ, критерий меньше порогового значения
1	0	401	318	0
2	0	265	304	0
3	69	132	226	0
4	0	249	152	26
5	0	260	287	0
6	0	176	205	43
7	0	136	305	5
8	0	160	179	0
9	0	149	220	0

Как видно из таблиц 8-13, существуют критерии, позволяющие производить попарное различение любых двух звуков из числа звуков З, ЗЬ, Ж, ЖЬ для всех респондентов одновременно, с надежностью, превышающей 50 процентов, а во многих случаях – и намного большей. Однако в некоторых случаях надежность лишь слегка превышает 50 процентов, что не может быть признано удовлетворительным. В этих случаях можно поступить следующим образом.

1. Представленные в таблице 7 критерии – это *лучшие* из критериев, позволяющие различать указанные пары звуков. Однако существуют и другие критерии, позволяющие делать то же самое, хотя и менее эффективно. К сожалению, объем статьи не позволяет привести их. Однако они есть, и их совместное применение с критериями из таблицы 7 позволяет повысить надежность распознавания.

2. Данные, приведенные в таблицах 8-13, получены на фрагментах звуковых кривых, имеющих длину около 3000 точек, что при частоте дискретизации 44100 Гц соответствует 0.068 секунды. Одновременное использование нескольких идущих подряд фрагментов повышает надежность распознавания звука.

3. Если достоверно известно, *кто* из респондентов произнес звук, то конкретно для него можно использовать и другие критерии. Так, если для респондента №2 использовать критерий I_2/I_7 при пороговом значении 16, то таблица надежности выглядит так

Табл. 14. Различение звуков Ж и ЖЬ для респондента №2. Использован критерий I2/I7, его пороговое значение – 36

Респондент	Звук Ж, критерий больше порогового значения	Звук Ж, критерий меньше порогового значения	Звук ЖЬ, критерий больше порогового значения	Звук ЖЬ, критерий меньше порогового значения
2	59	324	272	32

Сравнение таблицы 9 с таблицей 14 показывает, насколько увеличилась распознаваемость звука ЖЬ для этого респондента.

6. ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

В данной статье представлены несколько способов получения вероятностных критериев различения как звуков, произнесенных одним человеком, так и различных респондентов, произносящих один и тот же звук. Несомненно, что существуют и другие вероятностные критерии. Одновременное использование нескольких независимых вероятностных критериев повышает надежность распознавания до почти 100%. Что касается абсолютно надежных критериев, стопроцентных, то таких, скорее всего, нет и быть не может – распознавание носит принципиально вероятностный характер. Это подтверждается, во-первых, тем, что человек, слушающий другого человека, но не видящий его, иногда допускает ошибки распознавания как персоны, так и речи и, во-вторых, тем, что звуки речи содержат в себе существенную

неустойчивость параметров – см. рисунок 3. С другой стороны, если бы распознавание не носило вероятностный характер, то один человек не смог бы понимать речь другого человека при наличии посторонних шумов, при каких-то изменениях в гортани, при наличии сильной усталости и т.д. Любой выход за какие-то рамки тут же приводил бы к отказу понимания. На самом же деле, во всех сомнительных случаях человек, пусть даже неосознанно, использует лингвистический контроль, он обладает способностью «дорисовывать» звуки и даже отдельные слова, услышанные в условиях плохой слышимости, наличия шумов и т.д.

ЛИТЕРАТУРА

1. Галунов В.И., Лобанов Б.М., Загоруйко Н.Г. Синтез и распознавание речи (попытка построения онтологии) // Акустика речи: материалы 14-й сессии российского акустического общества. Н.Новгород, 15 – 18 ноября 2004.
2. Лобанов Б.М. О развитии речевых технологий в Беларуси. // Речевые технологии. − 2008. − №1. − с. 49 − 59.
3. Сорокин В.Н. Фундаментальные исследования речи и прикладные задачи речевых технологий. // Речевые технологии. − 2008. − №1 − с. 18 − 48.
4. http://intsys.msu.ru/invest/speech/research. Интеллектуальные системы. Сайт кафедры МТИС и лаборатории теоретической кибернетики механико-математического ф-та МГУ.
5. Бабин Д.Н., Мазуренко И.Л., Холоденко А.Б. О перспективах создания системы автоматического распознавания слитной устной русской речи. // Интеллектуальные системы. − 2004 − Т.8, № 1 − 4. − с. 45 − 70.
6. Лобанов Б.М., Цирульник Л.И. Компьютерный синтез и клонирование речи / − Мн. : Белорусская наука, 2008. − 342 с.
7. Гецевич Ю.С., Лобанов Б.М. Система синтеза белорусской речи по тексту. // Речевые технологии. − 2010. − №1. − с. 91 − 100.
8. Митянок В.В., Коновалова Н.В. Применение фазового анализа звуков речи для распознавания человека по его голосу. [Электронный ресурс] //Техническая акустика. – Электрон. журн. − 2013. − №4. − Режим доступа: http://www.ejta.org, свободный.
9. Митянок В.В. О числовых характеристиках некоторых низкочастотных звуков человеческой речи [Электронный ресурс] // Техническая акустика. – Электрон. журн. − 2008. − №15. − Режим доступа: http://www.ejta.org, свободный.
10. Митянок В.В. Определение числовых характеристик высокочастотных звуков речи на основе аппроксимации гармоническими функциями. // Известия НАН Беларуси, сер. ф.-м.н. − 2009. − №2 − с. 111 − 118.