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

2015, 7

В. В. Митянок

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

К проблеме идентификации и верификации личности по фазовым характеристикам звуков речи

Получена 14.07.2015, опубликована 27.07.2015

Метод аппроксимации использован для изучения гласных звуков речи человека. Показано, что в спектре гласных звуков присутствуют полуцелые (от базовой) несущие частоты. Обнаружены новые фазовые критерии различения людей по звукам речи. Показано, что опубликованные ранее фазовые критерии являются следствием новых, причем дисперсии новых критериев – систематически меньше, чем дисперсии прежних. Приводятся таблицы и графики, иллюстрирующие полученные результаты.

Ключевые слова: автоматическое распознавание речи, идентификация и верификация личности

ВВЕДЕНИЕ

Одной из задач современной акустики является задача компьютерной верификации и идентификации личности по голосу. Решение этой задачи, в частности, дало бы возможность идентифицировать личность на уровне, имеющем доказательную юридическую силу подобно тому, как в настоящее время имеет юридическую силу идентификация личности по отпечаткам пальцев. Однако, несмотря на значительные усилия, эта задача еще далека от завершения под «ключ». Так, в [1] отмечено, что «Распознавание лица, голос которого звучит на фонограмме, для целей уголовного судопроизводства может иметь только вероятностный характер». В [2] указано, что «специальный тест с парным сравнением речевых сигналов длительностью 5 секунд показал 53% правильного распознавания фонетистами». Значимость решения этой задачи, современное состояние вопроса, перспективы решения весьма полно изложены в [2].

В настоящем исследовании продолжается изучение фазовых закономерностей звуков речи человека, начатое в [3]. Как хорошо известно, ухо человека не реагирует на фазы отдельных мод, составляющих звук. Если бы это было не так, то человек, скорее всего, не смог бы распознавать речь, произносимую кем-то по телевизору, радиоприемнику, телефону, так как такие каналы связи вносят искажения в фазы. Надо полагать, что реагирование на фазы также усложнило бы понимание речи одного человека другим, или даже вообще сделало бы это невозможным. Однако из этого

никоим образом не следует, что в *произносимых* звуках нет никаких фазовых закономерностей – звуки ведь произносятся другими органами человека.

Как хорошо известно, звуковые функции гласных звуков представляют собой сумму синусоид с медленно (по сравнению с базовой частотой) меняющимися параметрами – амплитудами, частотами, фазами. В [3] было показано, что для гласных звуков существуют определенные комбинации между фазами мод, уникальные для звука и для диктора. Эти комбинации имеют вид

$$Z = \sum_i \varphi_i - \sum_j \varphi_j , \tag{1}$$

при условии

$$\sum i = \sum j . \tag{2}$$

Здесь φ_i, φ_j – фазы мод номер *i, j*. Это дает основания надеяться, что можно будет разработать систему идентификации и верификации личности по голосу, основанную на фазовых характеристиках звуков речи. С другой стороны, поскольку ухо человека не реагирует на фазы, то для злонамеренного звукоподражателя, каким бы талантливым он ни был, должно стать сюрпризом то, что его можно будет отличить от «правильного» носителя голоса.

Результаты [3] получены методом аппроксимации [4–5], выдвинутого как альтернатива методу преобразований Фурье. Метод аппроксимации применяется для разложения (почти) периодических функций и их сумм на отдельные моды и основан на функционале

$$S = \sum_{i=1}^{n} (y_i - b_{0i} - \sum_{k=1}^{l_1} a_{ki} \sin(\omega_k i) - \sum_{k=1}^{l_1} b_{ki} \cos(\omega_k i))^2 + \alpha \sum_{i=1}^{n-1} (b_{0,i} - b_{0,i+1})^2 + \alpha \sum_{k=1}^{l_1} \sum_{i=1}^{n-1} (b_{k,i} - b_{k,i+1})^2 + \alpha \sum_{k=1}^{l_1} \sum_{i=1}^{n-1} (a_{k,i} - a_{k,i+1})^2 , \tag{3}$$

где y_i – аппроксимируемая функция, заданная своими значениями в *n* равноотстоящих точках, а комбинация

$$b_{0i} + \sum_{k=1}^{l_1} a_{ki} \sin(\omega_k i) + \sum_{k=1}^{l_1} b_{ki} \cos(\omega_k i) , \qquad i=1..n \tag{4}$$

входящая в (3) – аппроксимирующая функция, a_{ki}, b_{ki} – дрейфующие амплитуды синус- и косинус- волн, b_{0i} – дрейфующее начало отсчета, l_1 – количество аппроксимирующих мод, ω_k – их несущие частоты.

Набор несущих частот ω_k в [3] назван ловящей сетью. Вычисляя частные производные функционала *S* по a_{ki} и b_{0i}, b_{ki} и приравнивая их нулю, получаем систему линейных алгебраических уравнений относительно дрейфующих амплитуд синус- и косинус- волн, и дрейфующего начала отсчета. Решив эту систему, можно затем найти фазы мод. Найденные решения можно также подставить в (4). Пары слагаемых (4),

соответствующих одной и той же частоте, образуют отдельные моды. Сумму всех мод и дрейфующего начала отсчета естественно назвать восстановленным звуком.

При графическом отображении изучаемых сигналов на экране монитора оказалось удобным сопоставлять величину i, использующуюся в (3) и (4), значению x-координаты пикселя. В настоящем исследовании масштаб частот выбран так, что значению частоты ω_k=1 соответствует физическая частота 44100/(2π)≈7019 Герц. В этом случае, при прорисовке на экране монитора синусоидального сигнала, отснятого при частоте дискретизации 44100 герц, изменение горизонтальной координаты на один пиксель соответствует изменению фазы на 1 радиан.

При проведении настоящего исследования использованы те же самые данные, от тех же респондентов, что и в [3].

1. ПРЕИМУЩЕСТВА МЕТОДА АППРОКСИМАЦИИ

По сравнению с методом преобразований Фурье метод аппроксимации имеет ряд преимуществ. В частности, находимые методом аппроксимации частоты не получают никакого уширения вообще, нет и фальшивых максимумов. С помощью метода аппроксимации были получены и принципиально новые результаты. Так, в [6] было показано, что для звуков «З», «ЗЬ», «Ж», «ЖЬ» существуют несущие частоты, равные 0.5, 1.5, 2.5 и т.д. от базовой. Это вызывает необходимость перенумерации частот – если частоты нумеровать в порядке возрастания, то базовой частотой теперь следует считать частоту моды номер 2, что и принято в настоящей статье. Новые же частоты можно назвать полуцелыми (от базовой). Интенсивности мод, соответствующих полуцелым частотам, в среднем намного уступают интенсивностям соседних мод. Поэтому они и не видны на спектрограмме Фурье.

В связи с тем, что для звуков «З», «ЗЬ», «Ж», «ЖЬ» выявлено существование полуцелых частот [6], возникает предположение о том, что то же самое может иметь место и для гласных звуков. Если это так, то при добавлении в ловящую сеть полуцелых частот, моды, соответствующие базовой и кратной ей частотам, станут вести себя проще за счет того, что за «вспышки» [6], ответственность за которые ранее вынужденно брали на себя базовая и кратные ей частоты, теперь будут отвечать полуцелые частоты. Это предположение подтвердилось (рис. 1, 2).

То же самое имеет место и для других звуков и респондентов. Следовательно, полуцелые частоты реально существуют и для гласных звуков.

Далее, метод аппроксимации позволил более точно определять спектр сигнала. Так, на рис. 3 представлена усредненная спектрограмма звука «У», респондент № 1.

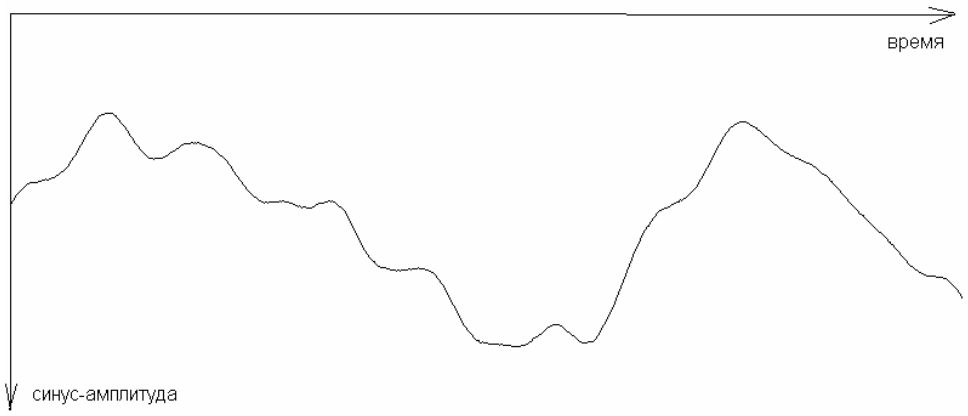

Рис. 1. Синус- амплитуда первой моды (ω=0.023) звука «А», респондент №1, полученная ловящей сетью, не учитывающей промежуточных частот

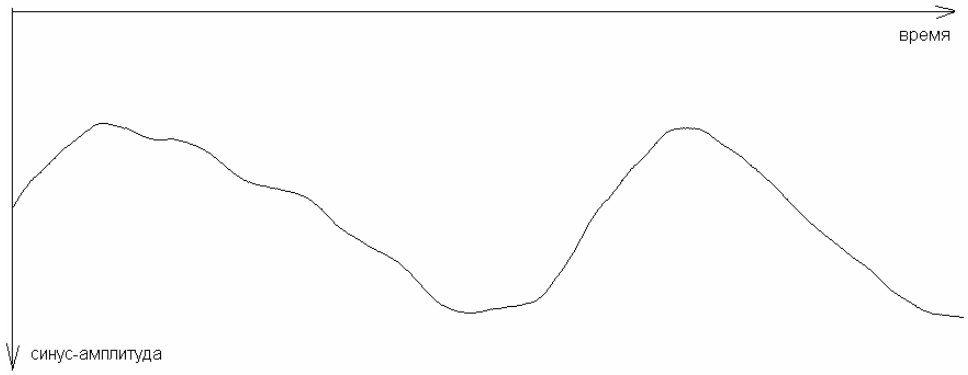

Рис. 2. Синус- амплитуда второй моды (ω=0.023) звука «А», респондент № 1, полученная ловящей сетью, учитывающей промежуточные частоты

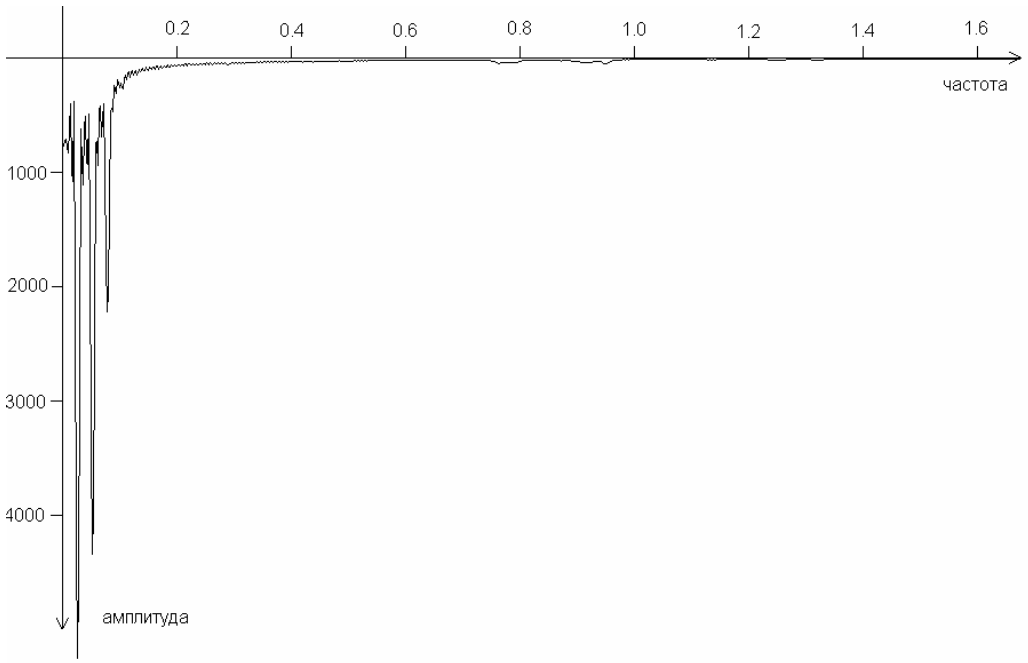

Рис. 3. Усредненная спектрограмма звука «У».

Для получения усредненной спектрограммы звуковая кривая звука «У» разрезалась на фрагменты длиной 1000 точек каждый, с перекрытием в 50%. Было получено около 200 фрагментов, вычислялись их спектрограммы, спектрограммы усреднялись по всем фрагментам. На усредненной спектрограмме (рис. 3) видны четкие линии, соответствующие частотам 0.026, 0.052, 0.078 (в наших единицах) и незначительное поднятие на участке от 0.7 до 1.6. Полуцелые частоты не видны. Чтобы разобраться с поднятием, масштаб по вертикали растянем в 100 раз (рис. 4).

На рис. 4 видны линии, соответствующие частотам 0.762, 0.791, 0.922, 0.950, 1.136, 1.250, 1.322, 1.575. Эти частоты, вместе с ранее найденными частотами 0.026, 0.052, 0.078 присутствуют в спектре звука «У».

Теперь составим ловящую сеть из 3-х частот, показанных на рис. 3, добавим к этим частотам полуцелые частоты, и проведем разложение изучаемого звука полученной сетью. После этого восстановим звук «У», и результат восстановления вычтем из исходного звука. То, что осталось после вычитания, подвергнем преобразованию Фурье (рис. 5).

В дополнение к частотам, обозначенным на рисунках 3 и 4, на рис.5 видны частоты 0.292, 0.501, 0.711, 1.036, 1.074, 1.426, 1.696. Все эти частоты, вместе с полуцелыми частотами 0.013, 0.039, 0.065 также присутствуют в спектре звука «У».

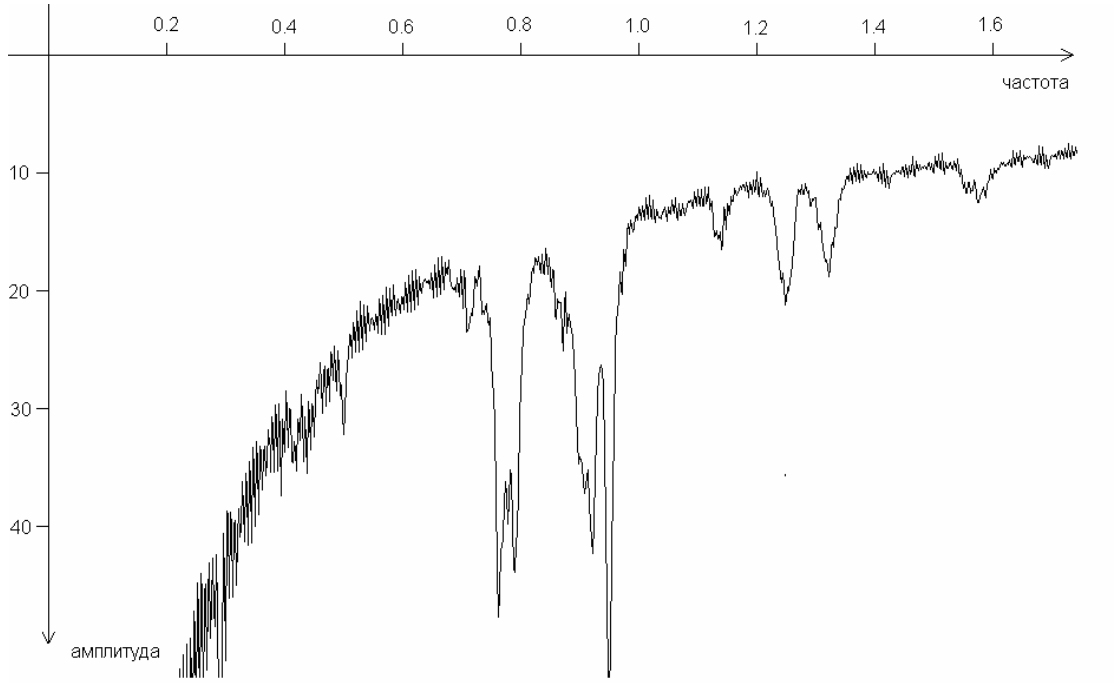

Рис. 4. То же самое, что и на рис. 3, только вертикальный масштаб растянут в 100 раз

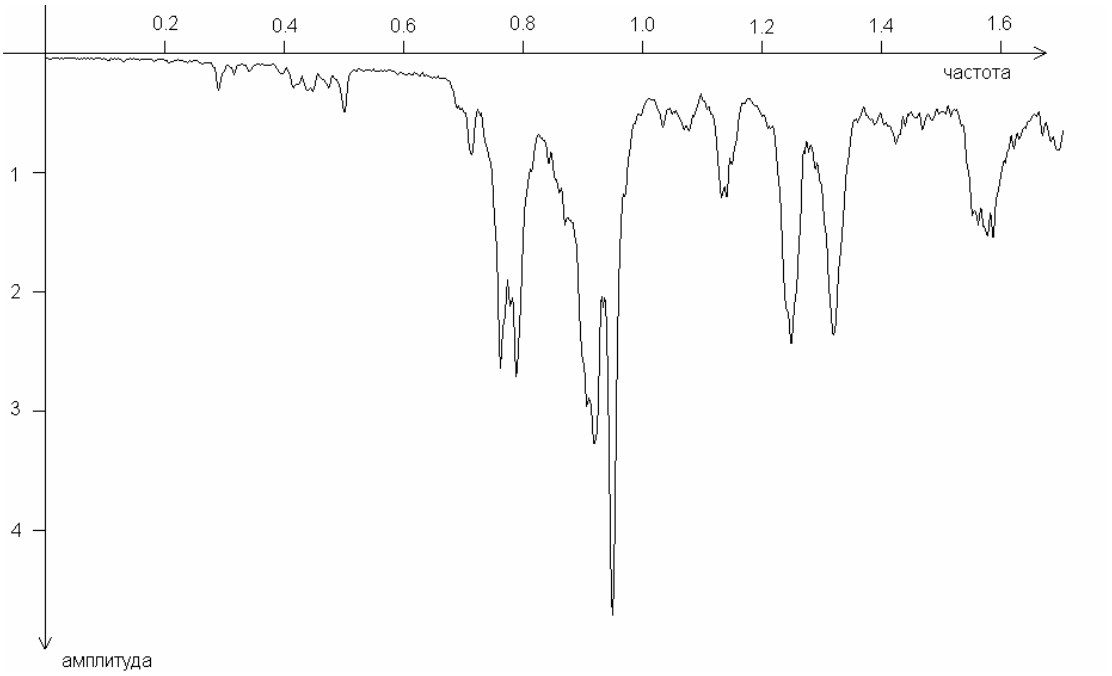

Рис. 5. Спектрограмма разности исходного звука «У» и восстановленного звука «У»

2. ФАЗОВЫЕ КОМБИНАЦИИ

Как выяснилось при проведении настоящего исследования, существуют еще более фундаментальные фазовые комбинации, нежели комбинации вида (1). Введем понятие нормированных фаз *четных* мод (для ловящей сети, содержащей полуцелые частоты). Они вычисляются по формуле

$$\varphi'_{2k} = \frac{\varphi_{2k}}{k}, \qquad k=1\ldots, \tag{5}$$

где $2k$ – номер моды, верхний штрих у значка фазы означает ее нормированность, целое число k в (5) может принимать значения в интервале 1-7. Бóльшие, чем 7, значения k нежелательны, так как с ростом номера моды возрастает значение ошибок в выборе базовой частоты, что может привести к скачкам фаз. Ниже приведем два примера поведения нормированных фаз в зависимости от времени. Нормированные фазы для других звуков и для других респондентов принципиально выглядят так же.

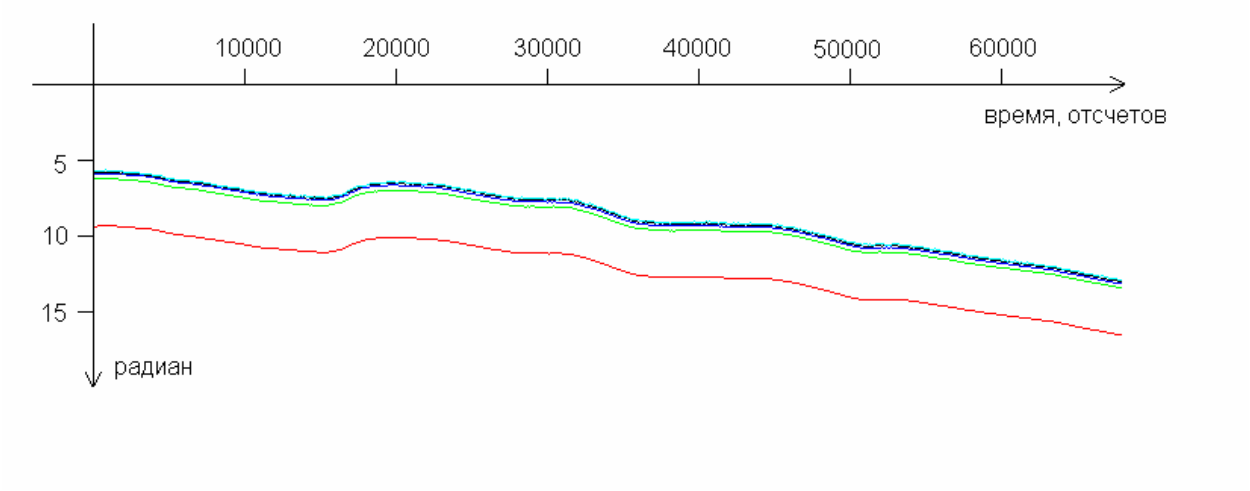

Рис. 6. Нормированные фазы четных мод. Респондент № 9, звук «Ы», базовая частота 274 Гц. Частота дискретизации 44100 Гц. Фазы разных мод имеют разный цвет

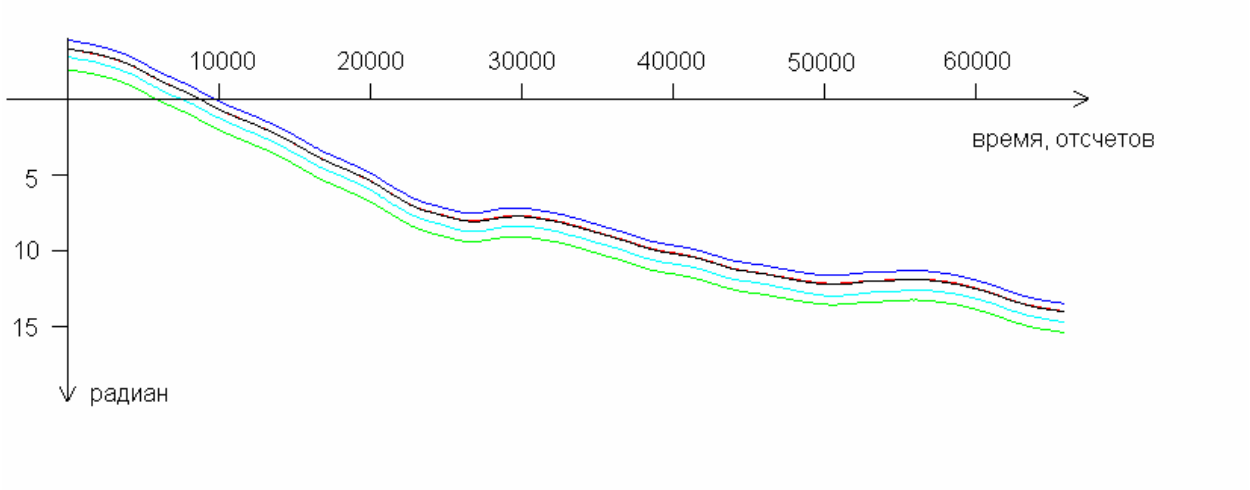

Рис. 7. Нормированные фазы четных мод. Респондент № 4, звук «Э», базовая частота 105 Гц. Частота дискретизации 44100 Гц. Фазы разных мод имеют разный цвет

Как видно из рисунков (6) и (7), нормированные фазы четных мод показывают одинаковую динамику. Расстояние между нормированными модами практически не зависит от времени. То есть, можно считать, что фаза одной из мод как бы управляет поведением остальных. Естественно предположить, что модой-водителем является базовая мода, то есть мода № 2.

Расстояния между различными нормированными фазами четных мод и фазой моды № 2 выражаются формулой

$$Y_k = \frac{\varphi_{2k}}{k} - \varphi_2, \qquad k=1..7. \tag{6}$$

Эти выражения и примем теперь в качестве новых критериев. Прежде всего, отметим, что фазовые критерии (1) и новые критерии (6) связаны между собой. Так, например, один из критериев (1) может быть представлен как линейная комбинация критериев (6):

$$\varphi_2 - 2\varphi_4 + \varphi_6 = 4(\varphi_2 - \frac{\varphi_4}{2}) - 3(\varphi_2 - \frac{\varphi_6}{3}). \quad (7)$$

Аналогичные разложения можно провести и для других критериев вида (1). Таким образом, критерии (1) являются следствием критериев (6), и, наоборот, критерии (6) являются более глубинным уровнем критериев (1).

Несложно показать, что критерии (6), также как и критерии (1), устойчивы по отношению к сдвигу начала времени и к небольшим ошибкам выбора базовой частоты. Доказательство легко проводится тем же способом, что и в [3].

Поскольку фазы являются периодическими величинами с периодом 2π, то критерии Y_k являются периодическими величинами с периодами $2\pi/k$. Из (6), в частности следует, что Y_1=0.

3. ОБРАБОТКА ОПЫТНЫХ ДАННЫХ

Звуковые кривые звуков «А», «О», «Э», «У», «Ы» полученных от 11 респондентов (использованы те же данные, что и в [3]), подвергались разложению на моды, вычислялись фазы мод, критерии Y_k, после чего критерии усреднялись по отрезкам длиной в 500 точек. Результаты усреднения приводились к интервалу $[0, 2\pi]$, после чего по этим данным составлялись гистограммы. Из-за ограниченности объема статьи приведем результаты лишь по одному из респондентов и лишь по одному из звуков (рис. 8).

Гистограммы критериев Y_2-Y_7 на рис. 8 разнесены по вертикали, начиная с критерия Y_2 (так как Y_1 все равно тождественно равен 0), сверху вниз, по очереди их номеров. Горизонтальный интервал $[0, 2\pi]$ растянут в 100 раз. Так как имеет место сокращенная периодичность критериев Y_k (по сравнению с 2π), то на каждой из горизонтальных осей появилось несколько группировок отличных от нуля значений гистограмм. Поскольку каждая из группировок оси описывает одну и ту же физическую ситуацию, то имеет смысл усреднить все группировки оси. Результаты усреднений были разнесены на рис. 8 по местам предыдущего нахождения группировок.

Приведем теперь совместные гистограммы для трех респондентов одновременно (рис. 9). В том же порядке, что и на рис. 8, гистограммы разных критериев разнесены по вертикали. Все респонденты, гистограммы которых представлены на рис. 9, –

женщины, с близкими базовыми частотами. Эти гистограммы могут быть использованы для различения этих респондентов по произносимому ими звуку А. Если какой-то образец критерия Y_2 имеет, например, значение 0.2, то звук был произнесен респондентом № 7, если этот критерий имеет значение 0.17, то звук был произнесен респондентом № 10, если этот критерий имеет значение 2.34, то звук был произнесен респондентом № 10. Ну, а если критерий имеет значение 0.3, то звук был произнесен кем угодно, кроме респондентов 6, 7, 10. На рис. 9 видно, что существуют зоны частичного перекрытия гистограмм. Так, если критерий Y_2 попал в интервал [1.89,1.92] то можно лишь утверждать, что звук был произнесен *не* респондентом 6, а вот различение респондентов 7 и 10 может быть проведено с лишь с некоторой вероятностью. Для более надежного их различения можно использовать критерии Y_3-Y_7 или другие звуки.

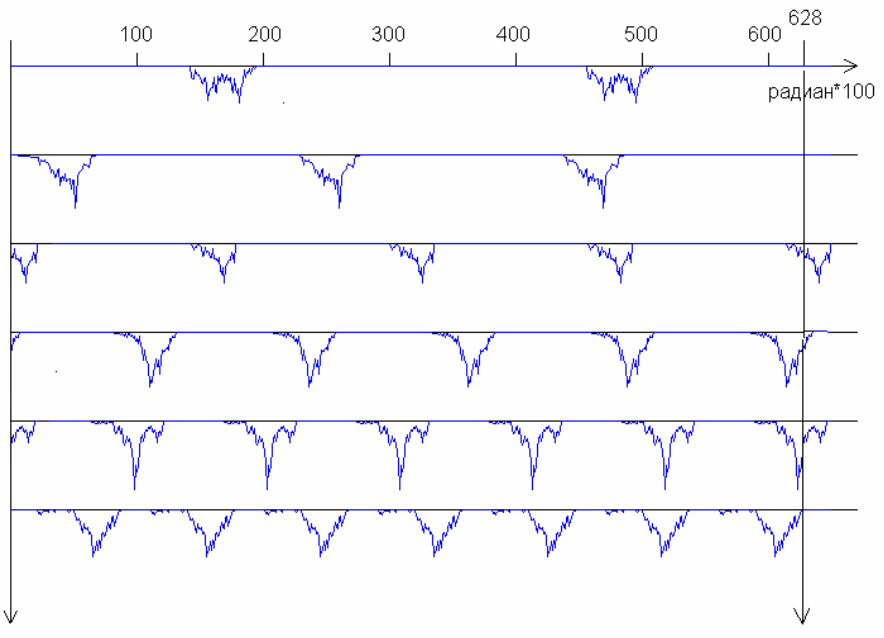

Рис. 8. Гистограммы фазовых критериев звука «А» в исполнении респондента №10

Приведем ниже таблицы числовых характеристик распределений критериев, полученных при анализе большого числа образцов звучания. Из-за ограниченности объема статьи данные приведены только по респонденту № 1 и только по звуку «А».

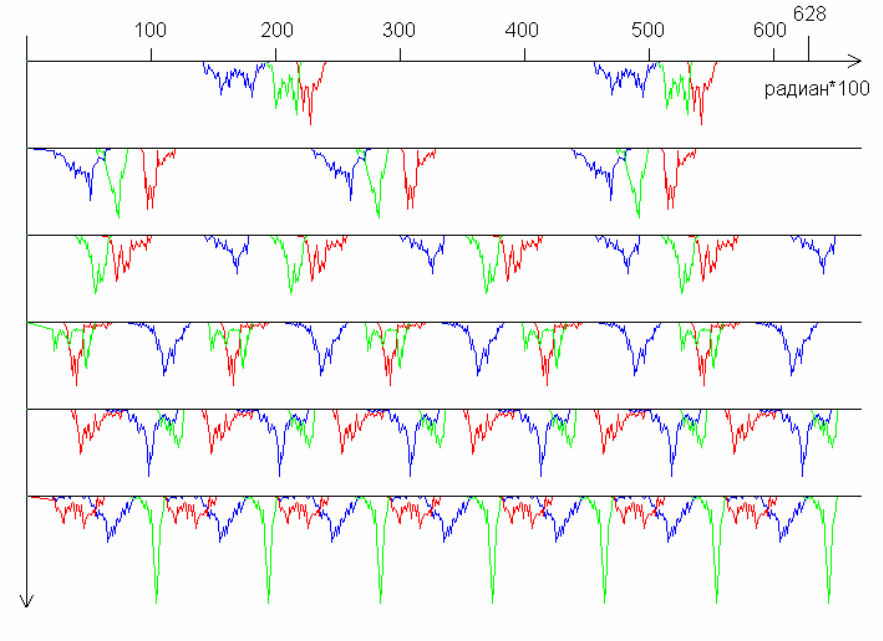

Рис.9. Гистограммы фазовых критериев звука «А» в исполнении респондентов № 6 – красные линии, № 7 – зеленые линии, № 10 – синие линии

Так как в процессе звучания базовая частота звука слегка менялась, то для разложения фрагментов звука на моды использовались несколько ловящих сетей. Всего в настоящем исследовании, для всех респондентов вместе взятых, использовались пропорциональные ловящие сети с базовыми частотами от 0.014 до 0.041 и с шагом 0.001. Каждый из фрагментов звука, длиной 1000 отсчетов, анализировался на предмет его базовой частоты, затем среди всех ловящих сетей выбиралась та, базовая частота которой была ближе всех к базовой частоте фрагмента, и по ней проводилось разложение. Затем вычислялись фазы, фазовые критерии, критерии усреднялись по каждому из фрагментов. Результат усреднения по фрагменту представлял собой как бы результат *одного* испытания. Затем вычислялись средние значения и среднеквадратические отклонения испытаний. Результаты приведены в нижеследующих таблицах. В первой строчке каждой из таблиц указываются критерии, во второй – их средние значения, домноженные на 100, в третьей – среднеквадратические отклонения, домноженные на 100, в четвертой – количество фрагментов, участвовавших в исследовании. Часто оказывалось так, что среди фактических значений критериев присутствовало несколько удаленных от общей группы значений. Они интерпретировались как выбросы и исключались. Поэтому числа четвертых строк несколько различны для разных столбцов. Всего количество выбросов составляло 0.5-2.0 процента от общего числа данных.

Табл. 1. Числовые характеристики распределения критериев полученных для звука «А». Базовая частота – 0.023

Критерий	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
среднее	64	79	137	11	65	11
ср.кв.от.	17.7	11.0	8.0	14.2	10.6	7.6
образцов	366	359	365	365	361	361

Табл. 2. Числовые характеристики распределения критериев полученных для звука «А». Базовая частота – 0.024

Критерий	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
среднее	52	85	149	18	80	17
ср.кв.от.	10.8	11.4	12.1	12.6	14.9	6.1
образцов	328	333	333	333	328	319

Табл. 3. Числовые характеристики распределения критериев полученных для звука «А». Базовая частота – 0.025

Критерий	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
среднее	47	95	162	30	92	25
ср.кв.от.	11.1	7.6	3.9	6.7	5.3	4.1
образцов	196	197	189	196	197	196

Табл. 4. Числовые характеристики распределения критериев полученных для звука «А». Все частоты вместе

Критерий	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
среднее	56	85	147	18	76	16
ср.кв.от.	15.9	12.2	13.0	14.2	15.9	8.5
образцов	896	890	896	896	892	887

ЗАКЛЮЧЕНИЕ

Так как фазовые критерии (1), обнаруженные в [3], являются линейными комбинациями фазовых критериев (6), то уже только из этого факта следует, что дисперсии критериев (6) - меньше, чем дисперсии критериев (1). Кроме того, дальнейшее уменьшение дисперсий оказалось возможным благодаря включению в ловящие сети полуцелых частот. Все это, вместе взятое, привело в итоге к уменьшению среднеквадратических отклонений критериев в 1.5 − 2.5 раза по сравнению с данными [3]. Таким образом, перспективы применения фазовых критериев для идентификации и верификации личностей по голосу становится все более реальными.

ЛИТЕРАТУРА

1. Муженская Н. Е. Экспертиза в российском законодательстве. Руководство-справочник для следователя, дознавателя, судьи. Изд-во Проспект, 2014 г. 744 стр.

2. Сорокин В. Н., Вьюгин В. В., Тананыкин А. А. Распознавание личности по голосу: аналитический обзор./ В. Н. Сорокин, В. В. Вьюгин, А. А. Тананыкин //Информационные процессы, – 2012. – Т12, – N.1 – С.1.

3. 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.

4. 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.

5. 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.

6. Митянок В. В. О физической структуре звуков З, ЗЬ, Ж, ЖЬ [Электронный ресурс] // Техническая акустика. – Электрон. журн. – 2014. – 9. – Режим доступа: http://www.ejta.org, свободный.