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

2013, 4

В. В. Митянок, Н. В. Коновалова

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

Применение фазового анализа звуков речи для распознавания человека по его голосу

Получена 02.04.2013, опубликована 21.05.2013

Метод аппроксимации используется для разложения различных звуков речи человека на составляющие их моды. Представлены данные о 5 конкретных звуках, полученных от 11 респондентов. Определена динамика амплитуд и фаз различных мод. Обнаружено, что фазы различных мод не являются независимыми случайными величинами, наоборот, между ними имеются зависимости, причем уникальные для каждого из респондентов. Это указывает на перспективу разработки компьютерной программы автоматической идентификации человека по его голосу на уровне, имеющем доказательную юридическую силу.

Ключевые слова: автоматическое распознавание речи, цифровая обработка сигналов, распознавание человека по голосу.

ВВЕДЕНИЕ

Задачи автоматического распознавания речи человека и автоматического распознавания говорящего по его голосу с одной стороны, близки, поскольку у них общий объект исследования, а с другой стороны в определенном смысле противоположны: в первом случае требуется распознать речь, независимо от того, *кто* говорит, во втором случае требуется распознать говорящего независимо от того, *что* он говорит. Несмотря на близость задач успехи в их решении существенно различны. Если к настоящему времени уже разработаны и используются компьютерные программы, пусть и далекие от совершенства, но все же более-менее успешно распознающие речь, то успехи в решении задачи распознавания говорящего намного более скромны.

История и современное состояние дел по этому вопросу весьма полно изложены в [1]. Там, в частности, на стр. 2 отмечено, что «согласно регулярным годовым отчетам Gartner Group лишь около 1% потенциальных покупателей удовлетворено эффективностью коммерческих систем распознавания диктора».

Определенный интерес для распознавания личности (или верификации ее) может представить использование биометрических данных [2], либо заранее задуманной парольной фразы [3]. Однако можно предвидеть, что если в этих направлениях и будет достигнут успех, то он будет временным — метод [2] основан на амплитудно-частотных характеристиках тела человека, то есть предполагает наличие не теле диктора каких-то датчиков, а метод [3] привязан к контрольной фразе. Между тем, всем

известно, что знакомые между собой люди легко узнают друг друга при разговоре по телефону. Без всяких датчиков и контрольных фраз.

Задачи автоматического распознавания человеческой речи и автоматического распознавания человека по его голосу на первый взгляд представляются не очень сложными. Часто отрезки кривых звукового давления являются периодическими (или почти периодическими) функциями времени. В этих случаях можно использовать преобразования Фурье. По результатам этих преобразований можно находить доверительные вероятности, доверительные интервалы для математических характеристик различных звуков, полученных от различных респондентов и тем самым можно попытаться найти способы различения отдельных звуков в составе речи и идентификации говорящего.

Однако этот, казалось бы, очевидный путь, привел лишь к частичному успеху. Более того, в последнее время прогресс в данном направлении явно замедлился. Скорее всего, это связано с тем, что те идеи, которые «лежали на поверхности», к настоящему времени выработаны, и для дальнейшего продвижения вперед необходимо привлекать новые. В связи с этим обратимся к методу аппроксимации — принципиально иному способу решения обозначенных проблем, предложенному в [4, 5].

1. МЕТОД АППРОКСИМАЦИИ

В [4, 5] предложен метод аппроксимации для разложения сигнала любого происхождения, представляющего собой сумму почти гармонических слагаемых с медленно меняющимися параметрами (дрейфующими амплитудами, частотами, фазами), на исходные составляющие. Там же этот метод применен к анализу отдельных звуков человеческой речи. Метод основан на функционале

$$S = \sum_{i=1}^{n} \left[y(t_i) - y_1(t_i) \right]^2 + \alpha \sum_{i=1}^{n-1} (b_{0,i} - b_{0,i+1})^2 + \alpha \sum_{k=1}^{l} \sum_{i=1}^{n-1} (b_{k,i} - b_{k,i+1})^2 + \alpha \sum_{k=1}^{l} \sum_{i=1}^{n-1} (a_{k,i} - a_{k,i+1})^2, \quad (1)$$

где $y(t_i)$ — зависящая от времени аппроксимируемая функция, описывающая сигнал,

$$y_1(t_i) = b_{0,i} + \sum_{k=1}^{l} a_{k,i} \sin(\omega_k t_i) + \sum_{k=1}^{l} b_{k,i} \cos(\omega_k t_i) \quad (2)$$

— аппроксимирующая функция, $b_{0,i}$ — дрейфующее начало отсчета, $a_{k,i}$, $b_{k,i}$ — дрейфующие амплитуды волн, n — количество оцифрованных точек на аппроксимируемой функции, t_i — моменты времени, в которые заданы (оцифрованы) значения аппроксимируемой функции, l — количество складываемых волн (мод) в аппроксимирующей функции. В [4, 5] для простоты принято $t_i = i$, хотя это и не обязательно. Параметр α в (1) позволяет сглаживать изменения амплитуд волн при переходе от точки к точке. Дрейф амплитуд и начала отсчета означает, что эти величины также являются функциями времени.

При проведении разложения используется набор частот ω_k, по которым производится разложение. Этот набор частот можно назвать ловящей сетью. Частоты ловящей сети могут быть какими угодно, необязательно кратными базовой (низшей) частоте. Однако изучение формы кривых звукового давления звуков А, О У, Э, Ы и некоторых других показывает, что реальный речевой сигнал во многих случаях является почти периодическим, так что оправдан выбор такой ловящей сети, в которой частоты пропорциональны низшей. Такую ловящую сеть естественно назвать пропорциональной.

Если аппроксимируемой функцией является функция, состоящая из суммы синусоид (мод) с произвольными, но заранее заданными параметрами – частотами, фазами и амплитудами, и если число мод и их частоты известны, то процедура минимизации (1), проведенная согласно [4, 5], выведет на решение, соответствующее минимально возможной невязке, то есть равной нулю.

Вместе с тем, параметры *реальных* звуков испытывают некоторые дрожания, отклонения от постоянных значений. Это, в свою очередь, означает, что само понятие периода уже не является строго определенным. Поэтому следует быть готовым к тому, что речевой сигнал и ловящая сеть имеют различные базовые частоты. К каким последствиям это может привести?

Рассмотрим сначала идеальный гармонический сигнал

$$y(t) = A\sin(\omega t + \varphi) \tag{3}$$

имеющий амплитуду A, фазу φ, и частоту ω, и пусть ловящая сеть состоит из единственной частоты, имеющей значение $\omega_1 \neq \omega$. Проведем очевидные преобразования

$$\begin{aligned} &A\sin(\omega t + \varphi) = A\sin(\omega_1 t + \varphi + \Delta\omega t) = A\cos(\Delta\omega t)\sin(\omega_1 t + \varphi) + \\ &A\sin(\Delta\omega t)\cos(\omega_1 t + \varphi), \end{aligned} \tag{4}$$

где $\Delta\omega = \omega - \omega_1$, а коэффициенты перед $\sin(\omega_1 t + \varphi)$ и $\cos(\omega_1 t + \varphi)$ - суть дрейфующие амплитуды.

Как видно из (4), в случае ошибки в выборе базовой частоты, можно ожидать, что метод аппроксимации даст для дрейфующих амплитуд периодические функции с частотами, равными разности частот изучаемого сигнала и ловящей сети.

Однако можно предвидеть, что другие характеристики дрейфующих амплитуд, полученных методом аппроксимации, будут все же несколько отличаться от амплитуд, представленных в (4). В самом деле, если в (3) и (4) конкретизировать моменты времени t_i, а затем аппроксимируемую функцию (3) и дрейфующие амплитуды предполагаемой аппроксимирующей функции (4) подставить в (1), то первое слагаемое (1) (не содержащее α) будет равно нулю, а следующие, ответственные за гладкость дрейфующих амплитуд нулю не равны, что противоречит главной идее метода наименьших квадратов, когда функционал невязки составляется как сумма взаимно-антагонистических слагаемых, и минимум функционала достигается лишь тогда, когда *каждое* из слагаемых «идет на уступки» другому. А это не есть рассматриваемый случай.

Как показали численные эксперименты, проведенные авторами с искусственными сигналами, представляющими собой сумму нескольких мод с различными, но постоянными параметрами, и в самом деле, в случае небольшого несовпадения базовых частот сигнала и пропорциональной ловящей сети, найденные дрейфующие синус- и косинус- амплитуды сигнала представляют собой периодические функции, имеющие определенную частоту. При небольших ошибках в выборе базовой частоты ловящей сети всякий раз полученная методом аппроксимации частота найденной дрейфующей амплитуды оказывалась равной разности истинной частоты изучаемой моды и соответствующей ей по номеру частоты ловящей сети, как на это указывает выражение (4). А вот по величине дрейфующие амплитуды несколько отличались от предполагаемого по (4) решения.

Таким образом, можно, изучив поведение найденных дрейфующих амплитуд, выйти на верное значение базовой частоты изучаемого сигнала, внести коррективы в используемую ловящую сеть, вновь провести процедуру аппроксимации и теперь уже найти также и верные амплитуды и фазы.

Все это верно в том случае, когда модули разностей всех частот ловящей сети и соответствующих им по номеру частот мод изучаемого сигнала не превосходят половины каждой из базовых частот (то есть половины расстояния между частотами соседних мод). Как показали численные эксперименты, в противном случае факт периодичности дрейфующих амплитуд сохраняется, но скачком может измениться их частота. Следовательно, определенные ограничения на погрешность выбора базовой частоты все же существуют.

Как хорошо известно, звуки речи человека представляют собой сумму мод с различными частотами, амплитудами, фазами. Среди специалистов, занимающихся проблемой автоматического распознавания речи, распространено мнение о том, что слуховой аппарат человека не воспринимает фазу звукового сигнала. Это, разумеется, не означает, что в произносимых звуках нет никаких закономерностей, связанных с фазами отдельных мод. В связи с этим выдвинем предположение, что такие закономерности могут существовать. В первую очередь это должно касаться комбинаций (ниже — критериев) вида

$$Z = \sum_i \varphi_i - \sum_k \varphi_k \qquad (5)$$

при условии, что сумма номеров мод, входящих в (5) со знаком плюс равна сумме номеров мод, входящих в (5) со знаком минус, то есть, если

$$\sum i = \sum k \,. \qquad (6)$$

В (5) i и k – номера мод звукового сигнала. Фазы, составляющие комбинации, могут входить в них неоднократно. Выбор комбинаций в виде (5) обоснован тем, что они обладают двумя видами устойчивости. Во-первых, эти комбинации не зависят от выбора начала отсчета времени, и, во-вторых, они не зависят от небольших погрешностей выбора базовой частоты пропорциональной ловящей сети.

В самом деле, пусть аппроксимируемая функция имеет вид

$$y(t) = b_0 + \sum_{k=1}^{l} a_k \sin(\omega_k t + \varphi_k). \quad (7)$$

Осуществим в (7) сдвиг начала отсчета времени преобразованием $t = t' + \Delta t$. Тогда (7) принимает вид

$$y(t') = b_0 + \sum_{k=1}^{l} a_k \sin(\omega_k t' + \omega_k \Delta t + \varphi_k) = b_0 + \sum_{k=1}^{l} a_k \sin(\omega_k t' + \varphi'_k). \quad (8)$$

Как видно из (8), фазы изменились, новые фазы связаны со старыми соотношениями

$$\varphi'_k = \varphi_k + \omega_k \Delta t = \varphi_k + k\omega_1 \Delta t, \quad k=1...l. \quad (9)$$

Здесь ω_1 — частота первой из мод, она же базовая частота. Подставив (9) в (5) несложно убедиться в том, что в силу (6) значение комбинации (5) не меняется.

Теперь рассмотрим ошибку выбора базовой частоты. Сдвиг базовой частоты $\omega_1 = \omega'_1 + \Delta\omega_1$ и пропорциональный сдвиг высших частот $\omega_k = \omega'_k + \Delta\omega_k = \omega'_k + k\Delta\omega_1$ в (7) приводит к следующим изменениям.

$$y(t) = b_0 + \sum_{k=1}^{l} a_k \sin(\omega'_k t + \Delta\omega_k t + \varphi_k) = b_0 + \sum_{k=1}^{l} a_k \sin(\omega'_k t + \varphi'_k) \quad (10)$$

где

$$\varphi'_k = \varphi_k + \Delta\omega_k t = \varphi_k + k\Delta\omega_1 t, \quad k=1...l. \quad (11)$$

— новые дрейфующие фазы. Подставляя (11) в (5), нетрудно убедиться в том, что благодаря условию (6) значение комбинации также не меняется.

2. ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Для изучения фазовых критериев были собраны образцы звучания звуков «А», «О», «Э», «У», «Ы», полученых от 11 респондентов: 5 мужчин и 6 женщин. Звуки, полученные от каждого из респондентов в несколько приемов так, чтобы общая продолжительность звучания каждого из них составляла 10…12 секунд, вводились в компьютер через бытовой микрофон. Частота дискретизации составляла 44100 Гц. После ввода в компьютер звуковые функции разрезались на отрезки длиной около 1000 точек. Каждый из отрезков перекрывался с последующим и предыдущим отрезками на ½ своей длины. На каждом из отрезков двумя различными способами определялась базовая частота. Во-первых, звуковая функция подвергалась интегральному преобразованию Фурье, полученный спектр анализировался на предмет определения базовой частоты. Во-вторых, на звуковых кривых осуществлялся отбор характерных точек и определялась повторяемость этих точек при изменении времени. Если оба метода давали близкие результаты, то базовая частота усреднялась по обоим методам и считалась найденной, в противном случае отрезок забраковывался. Число забракованных отрезков составило около 10 процентов общего их числа. (Всего от

каждого из респондентов для каждого из звуков таким способом было получено от 700 до 1100 отрезков.)

После этого методом аппроксимации [4, 5] на каждом из отрезков проводилось вычисление дрейфующих амплитуд и дрейфующего начала отсчета. Число мод было принято равным 24. После этого, с целью нивелирования краевых эффектов, производилась обрезка вычисленных амплитуд с каждого из краев отрезка на ¼ его общей длины. Таким образом, длина принятой во внимание части отрезка составляла около 500 точек. Затем вычислялись фазы мод, составлялись их комбинации вида (5), (всего было рассмотрено 214 различных комбинаций), и результаты комбинирования усреднялись по всему отрезку. Конкретные числовые значения результатов усреднения ниже будем называть фазовыми величинами.

Поскольку каждое из слагаемых, входящих в (5), является периодической величиной с периодом 2π, то и сами фазовые комбинации (5) — также периодичны с тем же периодом, соответственно, периодическими являются и фазовые величины.

Если бы фазовые величины имели равномерное распределение на отрезке $[0,2\pi]$, то их среднеквадратическое отклонение составляло бы $\pi/\sqrt{3}\approx 1.81$ [6]. Однако оказалось, что в большинстве случаев среднеквадратическое отклонение намного меньше. Приведем результаты по наиболее компактным для каждого из респондентов распределениям фазовых величин.

Таблица 1. Наиболее удачные звуки и комбинации фаз для каждого из 11 респондентов. Числа в столбцах 4-6 даны в радианах. Числа в столбце 4 приведены к интервалу $[0,2\pi]$

1	2	3	4	5	6	7
респон-дент	фазовая комбинация	звук	среднее значение фазовых величин	средне-квадратич. отклоне-ние	общее среднекв. отклонение	количе-ство ве-личин
1	$2\varphi_1-\varphi_2$	У	5.56	0.21	1.2	846
2	$2\varphi_1-\varphi_2$	О	3.5	0.2	1.18	973
3	$2\varphi_1-\varphi_2$	Э	3.76	0.08	0.94	907
4	$2\varphi_1-\varphi_2$	Э	3.51	0.08	0.8	818
5	$\varphi_1+\varphi_2-\varphi_3$	Э	4.77	0.06	0.59	989
6	$2\varphi_1-\varphi_2$	Ы	5.86	0.07	0.72	963
7	$\varphi_1+\varphi_2-\varphi_3$	Э	6.2	0.13	0.79	876
8	$\varphi_1+2\varphi_2-\varphi_5$	Ы	0.88	0.13	1.0	778
9	$\varphi_1+\varphi_2-\varphi_3$	Э	0.05	0.07	0.62	1013
10	$2\varphi_1-\varphi_2$	Э	3.67	0.14	1.07	1016
11	$\varphi_1-2\varphi_2+\varphi_3$	А	5.17	0.22	1.07	962

В шестом столбце представлены усредненные по 214 критериям и 5 звукам среднеквадратические отклонения фазовых величин. Как видно, они в 1.5-3 раза меньше значения 1.81, характерного для случайного равномерного распределения. Средние значения фазовых величин, представленные в четвертом столбце, показывают что для разных респондентов они - заметно различны. Это может быть использовано для решения проблемы идентификации человека по его голосу.

К сожалению, объем статьи не позволяет представить и другие фазовые комбинации и звуки, пусть и не столь удачные, как представленные в таблице 1, но все же достаточно хорошие с точки зрения малости среднеквадратического отклонения.

Так как звуки, произносимые различными респондентами, отличаются также и по тональности, то, с целью нахождения способов идентификации человека по его голосу имеет смысл рассматривать двумерные диаграммы, по горизонтальной оси которых откладывается базовая частота, а по вертикальной — фазовые величины. Представим некоторые из диаграмм на нижеследующих рисунках. Данные, полученные от одного и того же респондента представлены точками, имеющими один и тот же цвет. Каждая точка на диаграмме соответствует одному отрезку звуковой кривой, то есть представляет одно значение фазовой величины. Стрелочка указывает на группировку данных (точек) полученных от конкретного респондента. Цифра у подножия стрелочки означает порядковый номер респондента.

Как видно из рисунков 1-4, точки, полученные от каждого из респондентов, ложатся на диаграммах «базовая частота – критерий» «кучно», то есть можно говорить о группировках точек. При этом группировки точек, соответствующих одному и тому же звуку, одному и тому же критерию, но полученных от различных респондентов, во многих случаях (хотя и не всегда) находятся в разных местах. Это означает, что различные комбинации фаз и в самом деле могут использоваться в качестве критериев, позволяющих различать человека по его голосу.

Различия в расположении группировок точек, полученных от разных респондентов, обусловлены, скорее всего, различиями в элементах речевого аппарата. Это означает, что для каждого респондента можно построить набор диаграмм, на которых будут представлены области, соответствующие различным звукам и различным критериям. Данный набор диаграмм будет как бы «голосовым портретом» респондента.

На каждом из представленных рисунков отсутствуют группировки точек для некоторых из респондентов. Это связано либо с тем, что для данного фазового критерия и для данного звука площадь, занимаемая точками отсутствующих респондентов, непомерно велика по сравнению площадями, занятыми точками от других респондентов, либо по причине того, что области, занимаемые точками от различных респондентов, полностью или почти полностью перекрываются. В этих случаях для идентификации респондента следует использовать другие критерии и другие звуки. Так, например, на рисунке 2 отсутствуют точки от респондента номер 7 по причине того, что они накладываются на область точек от респондента номер 6. Для того чтобы сделать выбор между респондентами 6 и 7 можно использовать критерий $\varphi_1 - 2\varphi_2 + \varphi_3$ и звук «Э» (рисунок 3), либо критерий $2\varphi_1 + \varphi_2 - \varphi_4$ и звук «Ы» (рисунок 1).

Все это верно и в других случаях отсутствия на каком-то из рисунков данных по ком-то из респондентов. Но даже если области, занимаемые точками от двух различных респондентов перекрываются частично (например, области респондентов 6 и 11 на рисунке 4), то и в этом случае можно говорить о различении этих двух респондентов с какой-то вероятностью, а одновременное использование нескольких вероятностных критериев различия позволит повысить вероятность идентификации до почти 100 процентов.

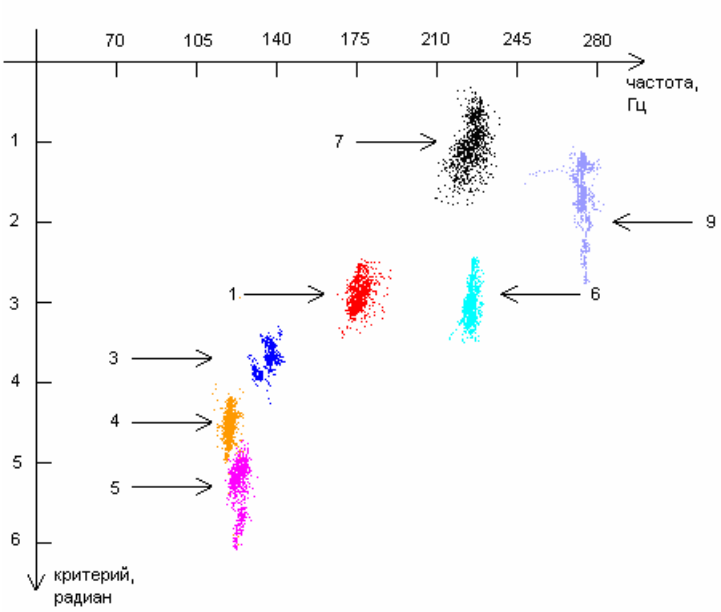

Рисунок 1.

Группировки точек для комбинации (критерия) $2\varphi_1 + \varphi_2 - \varphi_4$ и звука «Ы»

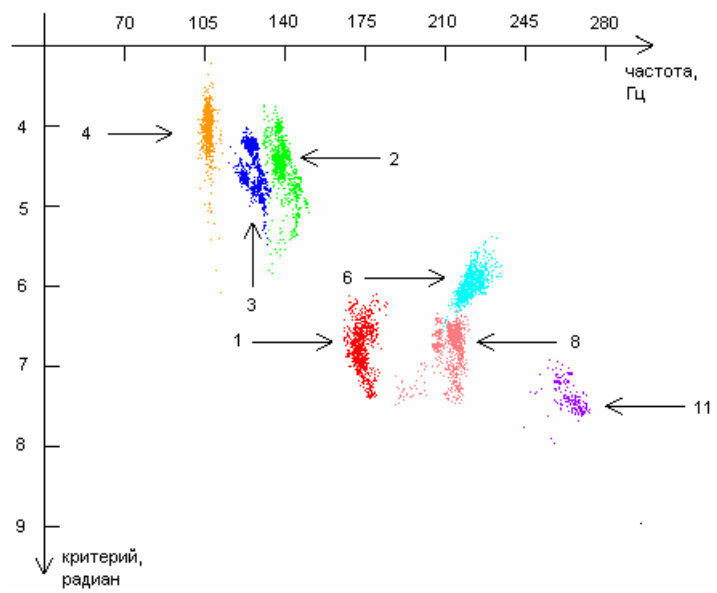

Рисунок 2.

Группировки точек для критерия $\varphi_1+\varphi_2 - \varphi_3$ и звука «О»

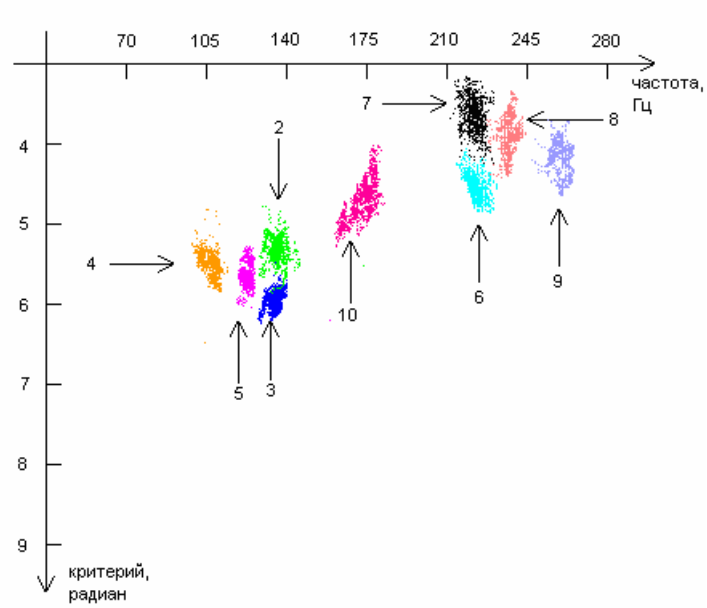

Рисунок 3.

Группировки точек для критерия $\varphi_1 - 2\varphi_2 + \varphi_3$ и звука «Э»

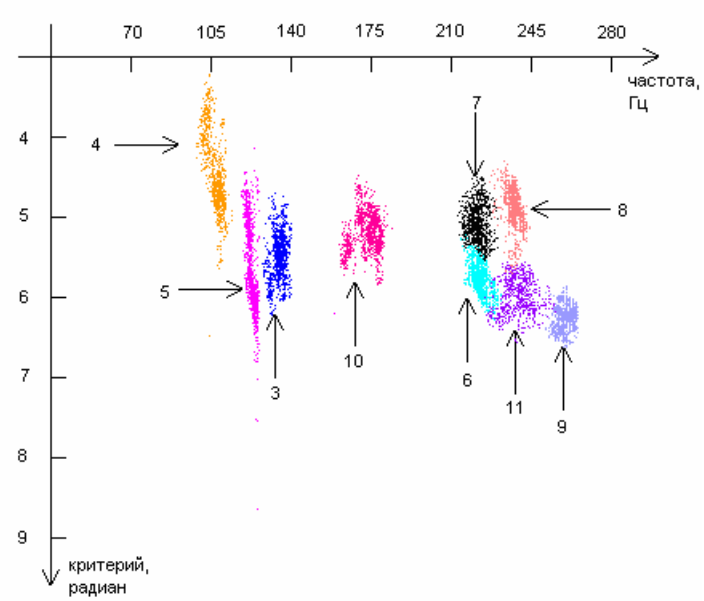

Рисунок 4.

Группировки точек для критерия $\varphi_1 - \varphi_2 - \varphi_5 + \varphi_6$ и звука «Э»

ЗАКЛЮЧЕНИЕ

В настоящее время для решения задач автоматического распознавания речи человека и автоматического распознавания говорящего по его голосу чаще всего используется метод преобразований Фурье. В то же время среди специалистов хорошо известно, что этот метод обладает рядом органических недостатков (см. например, [4, 5, 7]). Сам факт того что, несмотря на значительные усилия, указанные задачи до сих пор не решены, говорит о том, что здесь применение преобразований Фурье — скорее всего, ложный след.

С другой стороны альтернативный метод аппроксимации позволяет одновременно с амплитудами мод находить также и их фазы, что может быть использовано на практике. В нашем исследовании не оказалось ни одного случая, когда бы не нашлись,

(и даже неоднократно), фазовые критерии попарного различения 11 случайно отобранных респондентов. Некоторые из критериев позволяют одновременно различать по 7-9 респондентов. Это говорит о том, что найденные закономерности носят систематический характер, и что можно разработать компьютерную программу идентификации человека по его голосу на уровне, имеющем доказательную юридическую силу.

ЛИТЕРАТУРА

1. Сорокин, В.Н. Распознавание личности по голосу: аналитический обзор / В.Н. Сорокин, В.В. Вьюгин, А.А. Тананыкин // Информационные процессы. – 2012. – Т12. – N.1. – С.1.
2. Способ контактно-разностной акустической идентификации личности: Пат. РФ 2451346. МПК G10L17/00 / Дворянкин С.В., Голубинский А.Н. – N2011116633/08; заявл. 27.04.2011; опубл.20.05.2012 // Бюлл. N14 – 11 с.
3. Способ аутентификации диктора по парольной фразе: Пат РФ 2422920 и РФ 2422921. МПК G10L15/00 / Столов Е. Л. – заявка N2009106368/09; заявл.24.02.2009; опубл. 27.06.2011 // Бюлл. N18, – заявка N2009130688; заявл.11.08.2009; опубл. 27.06.2011; Бюлл. // N18.
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 / В.В. Митянок // Известия НАН Беларуси, Сер.ф.-м.н. – 2009. – N2. – С. 111.
6. Калинина, В.Н. Математическая статистика / В.Н. Калинина, В.Ф. Панкин. – М: Изд-во «Дрофа», 2002. – 336 с.
7. Воскобойников, Ю.Е. Фильтрация сигналов и изображений: Фурье и вейвлет алгоритмы / Ю.Е. Воскобойников, А.В. Гочаков, А.Б. Колкер. – Новосибирск: Изд-во «СИБСТРИН», 2010. – 195 с.