OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 534.87; 534.4

О ФИЗИЧЕСКОЙ СТРУКТУРЕ ПРОСТЫХ ГЛАСНЫХ ЗВУКОВ РЕЧИ ЧЕЛОВЕКА

Митянок В.В.

Полесский государственный университет
г. Пинск, Республика Беларусь

Mitsianok@mail.ru

В статье описываются результаты численных экспериментов по разложению гласных звуков речи человека на составляющие их звуковые моды с дрейфующими амплитудами. Полученные моды вновь суммировались, при этом предварительно подвергались сознательным искажениям с целью выявления факторов, как имеющих значение так и не имеющих значение для распознаваемости звуков.
Ключевые слова: распознавание речи, синтез речи, фазовый анализ звуков.

Введение

В последнее время появились сведения о том, что дельфины обладают разумной речью. [Janik и др., 2006; Kassewitz, 2015]**.** В частности, дельфины имеют имена, которые сородичи дают им при рождении, и по которым они обращаются друг к другу. Некоторые из звуков, издаваемых дельфинами, уже расшифрованы. Таким образом, в перспективе люди рано или поздно займутся масштабным изучением этой речи и, как следствие, ее семантикой. Однако в отношении дельфинов проблема семантики представляется намного более сложной, нежели в отношении человека, поскольку совершенно неизвестны ни мировоззрение дельфинов, ни их мышление. А оно, в свою очередь может быть понято лишь после установления контакта с ними. То есть возникает как бы замкнутый круг проблем. Проблемы усугубляются огромной разницей в частотных диапазонах речи человека и звуковых сигналов дельфинов – эти диапазоны перекрываются лишь частично, звуки дельфинов в основном лежат в ультразвуковой (по мнению человека) области частот.

Автор настоящей статьи разработал и демонстрирует на конференциях и научных семинарах компьютерную программу «перековки» голоса человека на другой частотный диапазон. Так, мужской голос может быть «перекован» на женский и даже детский, и наоборот.

Причем перековка происходит безо всякого ущерба для семантики. Это может быть шагом к унификации частотных диапазонов людей и дельфинов. Но унификация частотных диапазонов – это лишь часть дела. Необходимо выявить структуру основных звуковых единиц голосов дельфинов. А таковые в настоящее время не вполне ясны даже в отношении людей. Так, например, считается, что поскольку ухо человека не реагирует на фазы различных составляющих мод, то и в составе звуков, *издаваемых* человеком нет никаких фазовых закономерностей. Но это мнение было оспорено [Митянок и др., 2013].

Поэтому начинать нужно именно с выявления *математических* особенностей различных звуковых единиц как человека так и дельфинов. Представляет интерес вопрос о том, *что* именно делает звук «А» звуком «А», звук «О» звуком «О» и т.д. Какие именно математические характеристики звуков здесь существенны, какие привнесены несовершенством аппарата речеобразования человека, какие позволяют отличать одного диктора от другого, а какие вообще ни за что не несут ответственности, и попали в состав звуков случайно.

Расшифровка математических особенностей различных звуков речи человека – это ключ к расшифровке математических особенностей звуков речи дельфинов и, в более отдаленной перспективе, к пониманию их семантики.

1. Метод аппроксимации

Как известно, метод преобразований Фурье, используемый для нахождения спектра звуков, обладает рядом недостатков. В частности, в спектре присутствуют фальшивые линии, линии спектра даже в случае идеального гармонического сигнала, но рассмотренного на ограниченном

интервале времени, размыты. (В квантовой механике это обстоятельство является математической подоплекой соотношения неопределенностей). Спектр сигнала существенно зависит от его длительности. Если в исследуемом сигнале присутствуют малоинтенсивные моды, то они могут оказаться скрытыми под фальшивыми линиями. Поэтому в [Митянок, 2008, 2009] была поставлена задача нахождения спектра звуковых сигналов методом аппроксимации, с учетом того, что звуковые сигналы, соответствующие отдельным звукам речи человека представляют собой сумму мод, параметры которых (амплитуды, частоты, фазы) могут слегка меняться в процессе звучания, дрейфовать, дрожать, то есть зависеть от времени. Метод основан на функционале

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

где $y(t_i)$ — зависящая от времени аппроксимируемая функция, описывающая сигнал, заданная своими значениями в n последовательных моментах времени от t_1 до t_n ,*a*

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

— аппроксимирующая функция, $b_{0,i}$ — дрейфующее начало отсчета, $a_{k,i}$, $b_{k,i}$ — дрейфующие амплитуды синус- и косинус- волн (параметры аппроксимирующей функции), ω_k - их несущие частоты, l — количество волн (мод) в аппроксимирующей функции. В (1) и (2) для простоты можно принять $t_i = i$, хотя это и не обязательно. Параметр α в (1) позволяет сглаживать изменения амплитуд волн при переходе от точки к точке. Чем больше значение α, тем более гладкими являются амплитуды волн.

Вычисляя частные производные (1) по дрейфующим амплитудам и по дрейфующему началу отсчета и приравнивая результаты нулю, получим систему линейных алгебраических уравнений относительно параметров аппроксимирующей функции. Решив эту систему, найдем эти параметры и тем самым произведем разложение аппроксимируемой функции на сумму волн с медленно меняющимися амплитудами. Найденные таким путем $b_{0,i}$, $a_{k,i}$, $b_{k,i}$ можно вновь подставить в (2) и произвести численное суммирование. Полученную таким путем аппроксимирующую функцию естественно назвать восстановленным звуком. Если затем вычесть восстановленный звук из исходного звука и подвергнуть разность преобразованиям Фурье, то выясняется, что часто существуют еще какие-то несущие частоты, которые не были замечены при первом разложении в ряд (интеграл) Фурье по причине малой интенсивности несомых ими мод. В частности, этим способом в [Митянок, 2014] было установлено, что в спектре звуков «З», «ЗЬ», «Ж», «ЖЬ» присутствуют полуцелые (по отношению к базовой) несущие частоты.

Каждую из мод, входящую в (2) можно переписать в физически более информативном виде:

$$a_{k,i}\sin(\omega_k t_i) + b_{k,i}\cos(\omega_k t_i) = c_{k,i}\sin(\omega_k t_i + \varphi_{k,i}), \quad k=1..l, i=1..n \quad (3)$$

и тогда аппроксимирующая функция выглядит так:

$$y_1(t_i) = b_{0.i} + \sum_{k=1}^{l} c_{k,i}\sin(\omega_k t_i + \varphi_{k,i}). \quad i=1..n \quad (4)$$

Здесь $c_{k,i}$ - дрейфующая общая амплитуда моды, $\varphi_{k,i}$ - дрейфующая фаза моды.

2. Анализ простых гласных речи человека.

Изучались звуки «А», «О», «У», «Э», «Ы», «И». Эти звуки были отобраны для изучения по той причине, что их можно произносить достаточно долго и от этого они не теряют свою индивидуальность в отличие от звуков «Я», «Е» и других, которые при длительном звучании преобразуются соответственно в звуки «А», «Э» и т.д. Для изучения вышеуказанных звуков, методом преобразований Фурье определялась в нулевом приближении система несущих частот, нижняя из которых назначалась базовой. Во – вторых, эта система несущих частот дополнялась теми частотами, которые остались незамеченными методом Фурье при первом его использовании. Для этого использовался вышеописанный способ. В третьих, эта система частот дополнялась полуцелыми частотами, составляющими 0.5, 1.5, 2.5, 3.5 от базовой. В результате получалась ловящая сеть, которая и использовалась для окончательного разложения звуков на моды. При анализе разложенных звуков выявилось следующее:

Общие амплитуды различных мод заметно нестабильны в процессе звучания (рисунок.1). Их нестабильность носит хаотический характер, между дрейфующими (плавающими, болтающимися) амплитудами не прослеживается никакой связи.

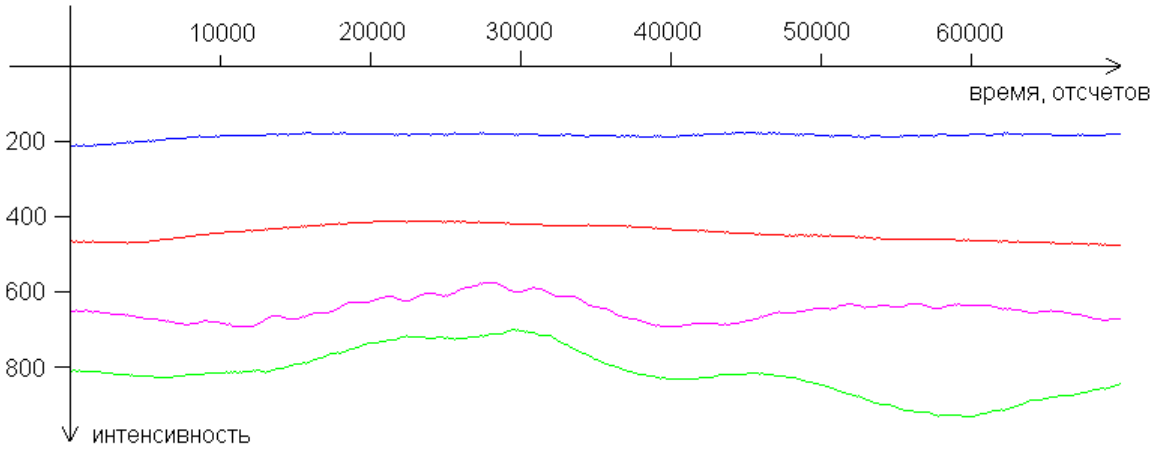

Рисунок 1. Амплитуды нижних мод, несомых целыми частотами. Звук «О», респондент Митянок. Частота дискретизации 44100 Гц. Базовая мода – красный цвет, вторая мода – фиолетовый, третья – зеленый, четвертая – синий. Образец N5

Напротив, между фазами отдельных мод такие связи прослеживаются достаточно хорошо (рисунки 2,3)

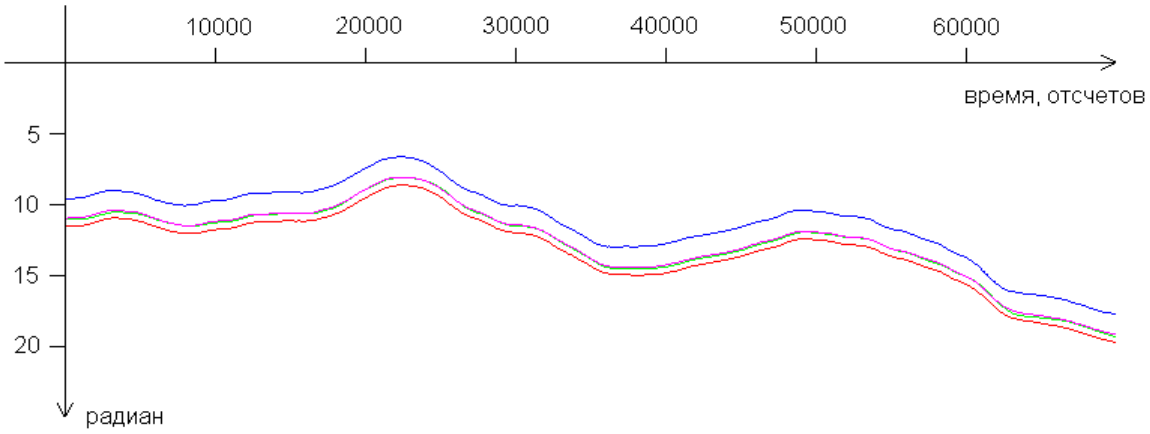

Рисунок 2. Фазы нижних мод, несомых целыми частотами, деленные на номер частоты (нормированные фазы).Звук «О», респондент Митянок. Частота дискретизации 44100 Гц. Базовая мода – красный цвет, вторая мода – фиолетовый, третья – зеленый, четвертая – синий. Образец N5.

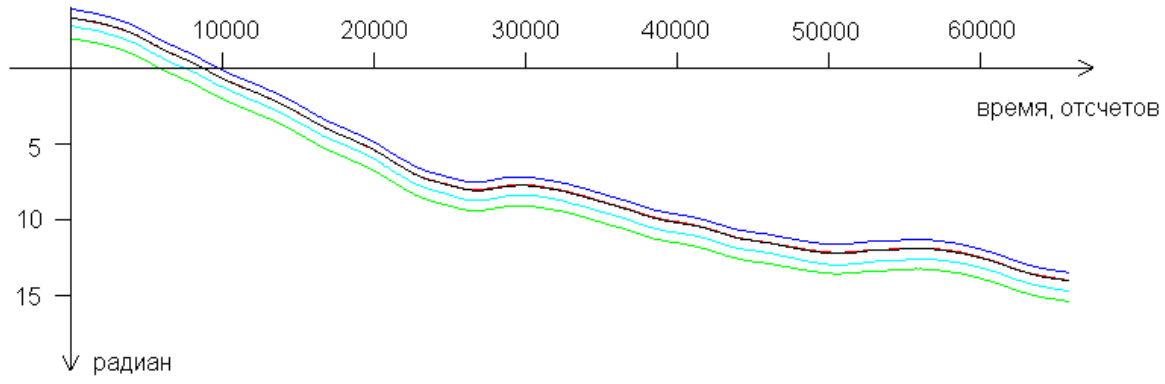

Рисунок. 3. Фазы мод, несомых целыми частотами, деленные на номер частоты (нормированные фазы). Звук «Э», респондент Янковский. Частота дискретизации 44100 Гц. Базовая мода – красный цвет, вторая мода – фиолетовый, третья – зеленый, четвертая – синий , пятая – бирюзовый.

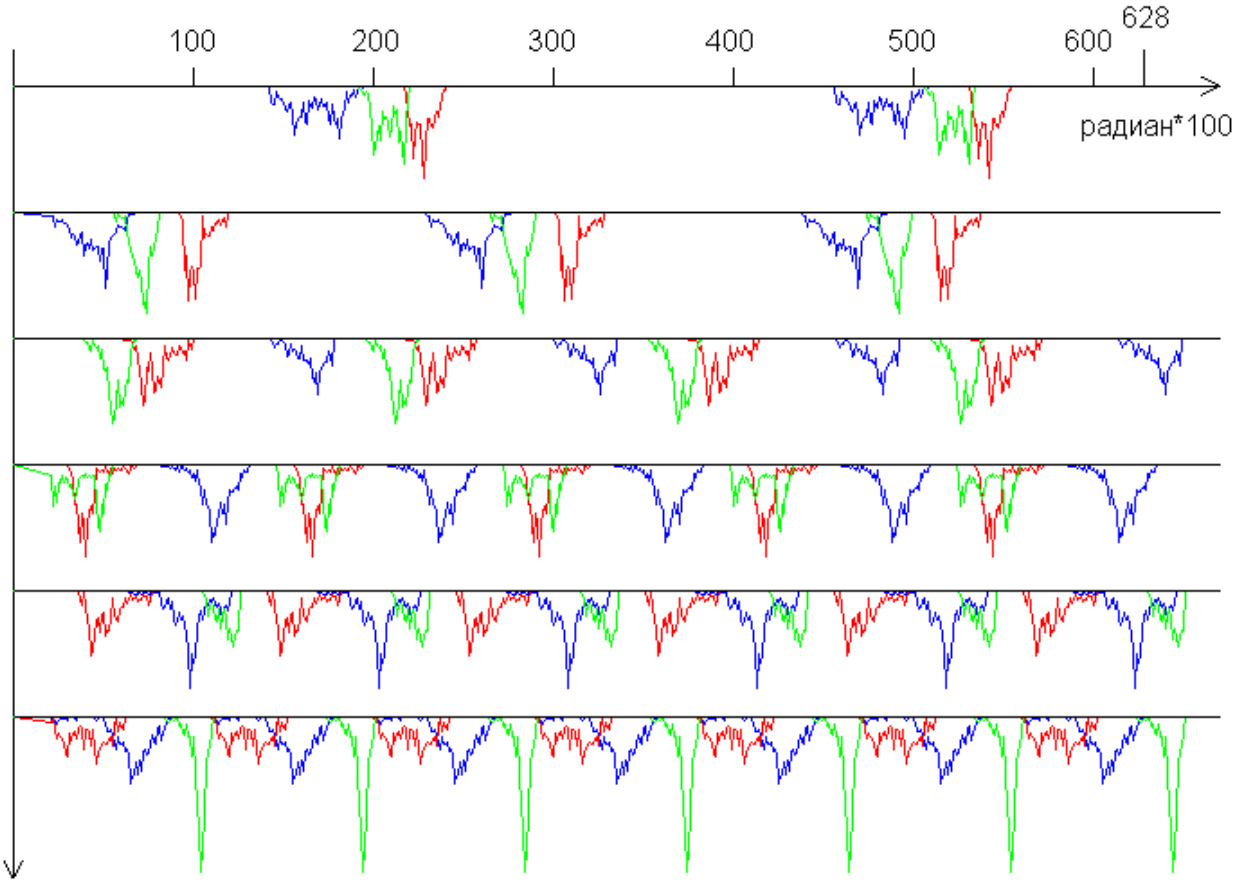

Рисунок 4. Гистограммы фазовых критериев звука А в исполнении респондентов: Коваленко – синие линии, Коновалова – красные, Романова – зеленые. Критерии разнесены по вертикали сверху вниз, начиная от K_2 и кончая K_7 Усреднение в (5) произведено при m=500. Частота дискретизации 44100 Гц

Как видно из рисунков 2 и 3 расстояние между различными фазами, нормированными на номер моды, практически не зависит от времени. Такое поведение фаз имеет место для всех исследованных звуков и для всех респондентов. Фаза моды, соответствующей базовой частоте может слегка зависеть от времени, нормированные же фазы остальных мод послушно повторяют ее изгибы. Однако расстояния между фазой базовой моды и нормированными фазами остальных мод, хотя и не зависят от времени, но зависят от того, какой именно звук звучит и от того, кто именно этот звук произносит. Это может быть использовано для создания систем идентификации и верификации человека по голосу (рисунок 4). Для этого введем фазовые критерии

$$K_i = \sum_{j=1}^{m} (\varphi_{1,j} - \frac{\varphi_{i,j}}{i}), i=1..l \qquad (5)$$

Здесь m –длина отрезка (в отсчетах дискретизации) избранная для проведения усреднения

При формировании гистограмм результат каждого усреднения рассматривался как одна точка. Периодичность критериев K_i из (5) составляет $2*\pi/i$, что объясняет тот факт, что в верхней строчке рисунка 4 на интервале $[0,2\pi]$ имеется 2 группировки точек, во второй строчке – 3, в третьей – 4 и т.д.

Уже только первая строка рисунка 4 позволяет по нескольким образцам звучания уверенно различать вышеуказанных трех респондентов. Пользуясь аналогичными гистограммами для других звуков речи, можно создать систему, позволяющими различать и намного большее число респондентов.

3. Синтез искусственных звуков

Если по формуле (4) произвести суммирование, используя найденные дрейфующие амплитуды волн, то, как и ожидается, получается звук, который при прослушивании звучит неотличимо от исходного звука. Для того, чтобы ответить на вопрос, *что* именно делает звук «А» звуком «А», звук «О» звуком «О» и т.д., перед суммированием были проведены математические эксперименты по сознательному искажению общих мод и фаз. Во-первых, фазы всех целых мод, кроме базовой, заменялись на искусственно вычисленные, привязанные к фазе базовой моды:

$$\varphi_{k,i} = k\varphi_{1,i}, \quad k=1..l, \; i=1..n \qquad (6)$$

где k – номер целочисленной моды, $\varphi_{1,i}$ - зависящая от времени (номера отсчета) фаза базовой моды. Звук от такой замены не меняется. Во- вторых, к новой фазе каждой из мод прибавлялось любое случайное число, постоянное, однако, на всем отрезке звучания. От этого звук также не менялся. В третьих, сам факт хаотичного поведения общих амплитуд различных мод различных звуков, полученных от различных респондентов и различных их образцов наводит на мысль о том, что хаотичность – это нечто привнесенное, не имеющее отношения к индивидуальности звуков. Так и оказалось. Оказалось, что дрейфующие общие амплитуды можно усреднить по отрезку звучания, и затем заменить фактические дрейфующие амплитуды их усредненными значениями. Звуковая функция, полученная после такого искажения амплитуд, звучала так же, как и исходный звук. Звук сохранял свою индивидуальность. В четвертых оказалось, что при суммировании мод можно опустить дрейфующий нуль и полуцелые частоты. И от такого отбрасывания звук не менялся. А вот если фазу каждой из мод на всем отрезке звучания заменить на постоянное, но случайное число, то звук портился. Вместо четкого звука слышалось то, что скорее можно назвать звучанием зуммера. В поисках объяснений были проделаны следующие математические эксперименты. 1. Усредненный амплитудный спектр любого из звуков соединялся в формуле (4) с фазами от любого из других звуков и от любого из других респондентов. От такой операции звук не менялся, звучал четко и соответствовал амплитудному спектру. 2.

Таблица 1. - Значения амплитуд различных мод простых гласных звуков.

Номер моды	А	О	У	Э	Ы	И
1	637	613	1060	566	1757	914
2	375	744	814	540	354	112
3	674	836	303	1007	65	22
4	794	495	0	61	0	0
5	753	51	0	114	25	0
6	180	0	0	123	51	0
7	49	0	0	90	140	0
8	19	0	0	97	32	0
9	15	0	0	183	54	16
10	17	0	28	93	111	49
11	17	10	0	114	14	30
12	21	17	0	120	10	35
13	8	0	0	44	22	71
14	16	0	34	31	92	135
15	16	0	7	42	26	147
16	16	15	17	54	26	110
17	30	30	8	79	8	35
18	34	12	0	45	0	6
19	13	0	0	37	0	8
20	0	0	0	18	7	5
21	0	0	0	45	0	14
22	0	0	0	25	9	21
23	0	12	0	0	6	14
24	0	0	0	0	12	5
25	0	0	10	0	10	9
26	0	0	10	0	8	22
27	0	0	5	0	7	16
28	0	0	3	0	0	24
29	0	11	4	0	0	42
30	0	15	8	0	0	18
31	0	18	13	0	0	14
32	0	18	5	0	0	13
Базовая частота	0.0269	0.0262	0.0305	0.0268	0.0302	0.0291

Примечание: допустимы небольшие (в пределах 10-30 процентов) изменения амплитуд, не влияющие на звук. Возможно также одновременное пропорциональное изменение всех амплитуд некоторого звука – этому соответствует изменение громкости. Данные получены усреднением по 20 образцам длительностью по 2-3 сек.

Перед пересадкой фаз с одного амплитудного спектра на другой, фаза базовой моды аппроксимировалась различными степенными многочленами степени от15 до 30. (при длительности звучания от 1 до 3 секунд). Результат был тот же – изменений в звучании нет. Так чем же объясняется «порча» звука при замене дрейфующих фаз на константы? Из рисунков 2,3 (и аналогичных для других звуков и для других респондентов) видно, что реальные фазы не являются строгими константами, а как бы дрейфуют (плавают) вокруг неких средних значений с неустойчивым периодом от 1.5 до 2.5 Гц и с неустойчивой амплитудой 0.5-2 радиан. В связи с этим возникло предположение, что именно так и должно быть. Что мозг слушателя уже готов к тому, что диктор будет производить сигнал с испорченной фазой, а звук с неиспорченной фазой мозгом за звук не воспринимается. Когда же в качестве фазы принималась испорченная величина, то звук вновь звучал четко и распознаваемо. Если подытожить все вышесказанное, то получаем, что для синтеза вышеуказанных звуков, вместо (4), как один из вариантов, можно принять формулу:

$$y_i = \sum_{j=1}^{32} C_j \sin\left[\omega_0 ij + 1.0\, j \sin(i/3300) + r_j\right], \quad i=1..n \qquad (7)$$

где усредненные значения общих амплитуд C_k приведены в нижеследующей таблице, ω_0 – базовая частота, ее значение приведено в последней строчке таблицы, r_k – массив произвольных чисел, n – длина отрезка звучания (в отсчетах дискретизации). За основу получения усредненных общих амплитуд в таблице 1 был взят голос автора. Изменению i в (7) на единицу соответствует изменение реального времени на 1/44100 долю секунды.

Заключение

В результате проведенных математических экспериментов установлено: 1. Звук определяется именно набором амплитуд, которые могут иметь постоянные значения по всему промежутку звучания. Возможные номинальные значения амплитуд приведены в таблице 1. 2. Моды частот, полуцелых по отношению к базовой, и дрейфующее начало отсчета несущественны для звуков. 3. Фазы мод, нормированных на номер моды, отличаются от фазы базовой моды на величину, постоянную по всему промежутку звучания, но зависящую от диктора и от произносимого им звука. 4. Прибавка к фазам мод произвольных постоянных чисел не влияет на звучание звуков. Заинтересованный читатель, пользуясь формулой (7) и данными для нее из таблицы (1) сможет самостоятельно подготовить любые из гласных звуков, исследованных в настоящей статье

Библиографический список

[Janik и др., 2006] Janik, V. Signature whistle conveys identity information of Bottlenose Dolphins/ V.M Janik, L.S. Sayigh, R.S. Wells // Proc. of the Nat. Acad. of Sci. of the USA 103, 21, 2006, pp 8293 – 8297.

[Kassewitz, 2015] Kassewitz, J. Speak Dolphin, 2015 -28 p.

[Митянок и др., 2013] Митянок , В В . Применение фазового анализа звуков речи для распознавания человека по его голосу. [Электронный ресурс] / В.В. Митянок , Н.В. Коновалова //Техническая акустика. – Электрон. журн.- 2013.-4.- Режим доступа: http://www.ejta.org, свободный.

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

[Митянок , 2008] Митянок,В.В. О числовых характеристиках некоторых низкочастотных звуков человеческой речи [Электронный ресурс] /В.В.Митянок // Техническая акустика. – Электрон. журн.- 2008.-15.- Режим доступа: http://www.ejta.org, свободный

[Митянок , 2009] 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.- , №2- с.111-118.

ABOUT THE PHYSICAL STRUCTURE OF SIMPLE VOWEL SOUNDS OF HUMAN SPEECH

Mitsianok V.V.

Polessian St. Univ
Pinsk, Belarus

Mitsianok@mail.ru

Approximation method is used for decomposition of the simple vowel sounds of human speech onto the set of the different frequencies partial waves, and for creating the artificial sounds. It is found, that before summation all the modes, their amplitudes may be averaged over the entire duration of sounds, so they may be constant values. The table of appropriate amplitude figures is presented. But appropriate phases should not be constants, but a bit spoiled constants. For regular phases sound sounds unnatural, but for slightly damaged phases the sounds are natural. The formula for synthesis of sounds is given. An interested reader can prepare artificial sounds by using this formula and table of averaged amplitudes.