УДК 004.934.2

ВЫДЕЛЕНИЕ ФОНЕМ ИЗ СЛИТНОЙ РЕЧИ И ИХ ИДЕНТИФИКАЦИЯ

Лелейтнер В.О.
АО НПП «Полигон», Уфа, РФ
E-mail: lel@ufacom.ru

Предложен способ определения местоположения фонологических формант, не связанный с анализом амплитудного спектра и обеспечивающий их выделение в слитной речи. Выделенные по данному признаку форманты группируются в фонемы, соответствие которых конкретному звуку определяется по минимальной дистанции от базовых фонем. Составлена таблица базовых фонем гласных и нескольких согласных звуков. Выявлены закономерность расположения базовых фонем и их привязка к координатам расположения на основной мембране кортиевого органа. Результаты работы предназначены для создания автономных автоматизированных систем распознавания речи.

Ключевые слова: *слитная речь, распознавание речи, фонема, форманта, кортиевый орган*

Введение

Несмотря на высокую скорость развития вычислительной техники и информационных технологий, основные проблемы речевых приложений до сих пор остаются актуальными. Основной причиной существующих проблем в распознавании речи является видимая сложность структуры речевого сигнала, содержащего огромное разнообразие фонетических единиц языка, интонационных окрасок и личностных особенностей говорящего. В результате речевые сигналы достаточно сложно детально исследовать и описывать с помощью математических моделей. Показательным является фактическое отсутствие систем распознавания русской речи со сверхбольшим словарем [1].

Наименьшим элементом речи является звук, который, как правило, в изолированном виде не существует. Точного определения понятия звука речи нет. Его, скорее всего, можно сравнивать с рукописной буквой [2]. Типизированные звуки речи в технике связи называются фонемами. Фонема – наименьшая звуковая единица данного языка, дифференцирующая слова и их формы и существующая в речи в целом ряде конкретных звуков – оттенков. Реализация фонемы, ее вариант, обусловленный конкретным фонетическим окружением, назван аллофоном.

Речевой сигнал представляет реакцию резонансной системы голосового тракта на возбуждение его одним или несколькими генераторами звуковых колебаний. Основные резонаторы образуются полостями рта и глотки, а в ряде случаев и носовой полостью. Области концентрации энергии в спектре звука речи, образуемые в том числе и основными резонаторами, называются формантами. Форманта, определяющая восприятие конкретного звука речи, называется фонологической формантой [2]. Кинематика речевого тракта в большинстве случаев позволяет произвести не более трех локальных сужений одновременно – на губах, кончике языка и в районе нёбной занавески. Это дает основания утверждать, что смысловая информация в речевом сигнале передается (для вокализованных звуков) параметрами первых трех формант [3].

Однако выделение фонологических формант вызывает значительную трудность, так как речевой тракт представляет собой многорезонансную систему, поэтому временной сигнал на его выходе есть результат наложения большого числа затухающих гармонических колебаний, а спектр амплитуд характеризуется множеством максимумов, которые являются ложными формантами, нехарактерными для данной фонемы [2; 4]. Кроме того, формантный максимум может раздваиваться, ложная форманта может иметь уровень выше основной [2]. Дополнительно спектральные составляющие основного тона часто маскируют первую форманту.

Поиски признаков и выделение инвариантных к диктору и контексту фонем продолжались вплоть до 90-х годов, но успеха не имели в том смысле, что ни одна из существующих в настоящее время систем распознавания речи результаты этих изысканий не использует. Отсутствие успеха в поиске локализованных во времени фонем объясняют тем фактом, что в естественной речи органы речеобразования практически никогда не занимают положений, характерных для изолированно произнесенных звуков, а лишь обозначают движение в нужном направлении, то есть речевой аппарат готовится к произнесению некоторых звуков заранее. Этот эффект называется коартикуляцией. Взаимовлияние фонем не ограничивается соседями, а может распространяться на несколько соседних фонем и даже на целое слово. В связи с этим, «используя аналогию с атомами, а лучше с квантами, можно заметить,

что фонема скорее имеет “волновую” природу, то есть ее признаки “размазаны” по протяженному во времени отрезку, причем признаки различных фонем накладываются друг на друга» [5]. Данные факторы приводят к отсутствию в общем случае соответствия фонетических символов и спектральных распределений, что доказано различными опытами и исследованиями. Между спектральной и фонетической функциями может быть установлено однозначное соответствие только при строгой стабилизации акустических условий и одном дикторе [6].

Для выбора произнесенной диктором фонемы используется многоступенчатая обработка на этапах предварительного выделения группы аллофонов, распознавания слов и морфем (наименьшая единица языка, имеющая некоторый смысл), лексического и смыслового контроля. На каждом этапе выполняется сложная обработка с использованием нейронных сетей, динамического программирования, скрытых и неоднородных марковских моделей и других методов. Целью методов обработки является нахождение имеющегося в базе данных образа, наиболее близкого к анализируемому образу фонемы, морфемы, слова и предложения.

Многообразие спектров фонем в слитной речи, сложность выделения фонологических формант и большие успехи в создании полосных вокодеров привели российских исследователей Варшавского Л.А. и Литвака И.М. к гипотезе о том, что фонетическое качество звуков определяется уровнем соотношений мощности в спектральных полосах, а форманты являются лишь доступным для речеобразующего аппарата способом достижения необходимых полосных соотношений. В начале 60-х годов была сформирована на основе большого экспериментального материала теория расчета разборчивости речи, принявшая за основу полосное представление речевого сигнала, исключавшая из рассмотрения форманты [7].

«Тонотопическая организация» периферической слуховой системы, при которой информация о спектральных компонентах, выделенная улиткой, проходит до соответствующих отделов центральной нервной системы, не перемешиваясь, принята за доказательство того, что амплитудный спектр сигнала является основой для распознавания речи человеком и, следовательно, для автоматических систем распознавания речи [5]. В связи с этими факторами в настоящее время в подавляющем большинстве систем распознавания для последующей обработки используется преобразование временного электрического сигнала в спектр Фурье.

В то же время положению о «волновой» природе фонемы противоречат результаты испытаний, в которых «несмотря на огромное разнообразие артикуляционных движений в связной речи и непрерывный характер речевых сигналов, говорящие на данном языке способны субъективно расчленять речь на фонемы. Фонетисты дают транскрипцию связной речи, используя разработанные для этой цели фонетические алфавиты» [4]. Кроме того, многочисленные видеограммы фраз связной речи [2, см. рисунки 6.25 и 10.30] показывают наличие достаточно выраженных границ между фонемами. Прослушивание слитной речи по частям также показывает, что органы слуха распознают звук речи в самом начале произношения звука, когда инструментальные характеристики не выделяют достоверные признаки фонемы. Проведенные автором эксперименты на гласных звуках показали, что идентификация звука происходит при длительности отрезка не менее 10…12 мс на всем протяжении звука по видеограмме.

Так как основу распознавания звуков речи человеком и, следовательно, автоматическими системами распознавания речи составляет амплитудный спектр, не акцентируется внимание на том факте, что улитка, являющаяся главным элементом периферической слуховой системы, не обладая сильными резонансными свойствами, представляет скорее линию задержки или временной анализатор [5].

Известно, что под воздействием входного звукового сигнала в улитке возникают две бегущие волны. Одна волна возникает в основной мембране, скорость распространения которой вдоль мембраны равна 50 мм/мс в непосредственной близости от овального окна, и, уменьшаясь по экспоненциальному закону, достигает у геликотремы значения 1,5 мм/мс. Скорость другой звуковой волны, распространяемой в перилимфе, в среднем равна 1500 мм/мсек. В связи с этим логично предположить, что на чувствительные клетки воздействуют два сигнала, при этом в каждой точке мембраны имеется различный временной сдвиг между воздействующими сигналами [8]. В зависимости от характера взаимодействия элементов кортиевого органа на чувствительные клетки возможно суммарное или разностное воздействие сигналов. При разностном воздействии в точке мембраны, для которой задержка между сигналами равна периоду определенной частоты, происходит частичная компенсация сигнала данной частоты и, соответственно, уровня общего сигнала. При суммарном воздействии частичная

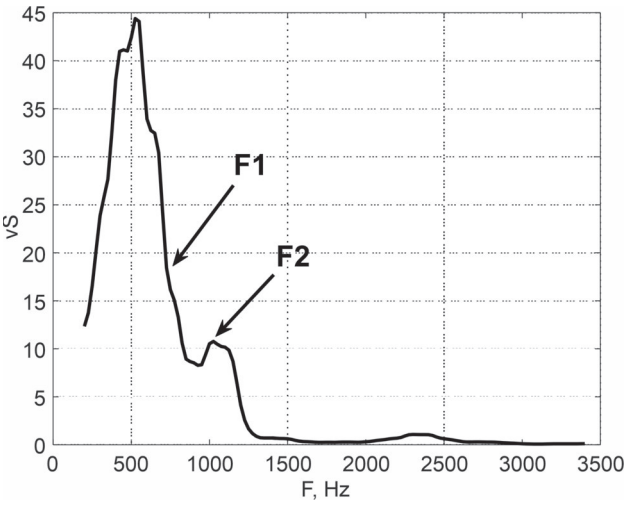

Рисунок 1. Спектр звука «а» из слога «АО»

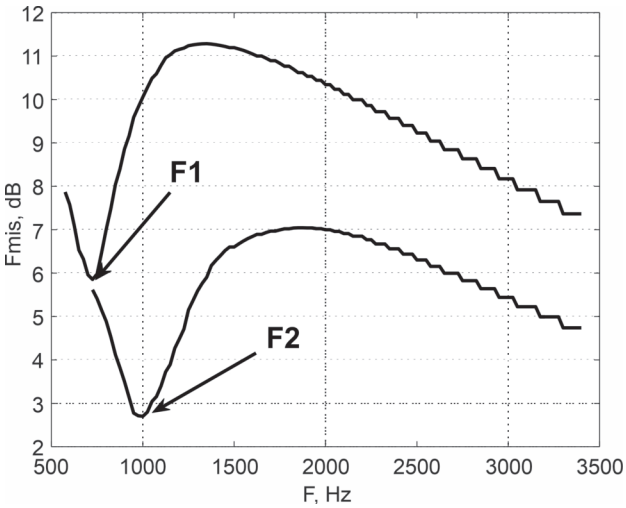

Рисунок 2. График функций суммарных сигналов первой и второй формант звука «А» из слога «АО»

компенсация происходит в точке мембраны, для которой задержка между сигналами равна полупериоду частоты. Вполне вероятно, что, используя данный механизм, слуховой аппарат выделяет из спектра частот частотные группы, которые являются фонологическими формантами.

Целью исследования стало определение возможности выделения фонологических формант в слитной речи на основе сложения или вычитания речевых сигналов с различным временным сдвигом и выявление закономерностей в организации системы распознавания звуков речи.

Результаты экспериментов

Определение возможности выделения фонологических формант выполнялось моделированием преобразований в среде Matlab. Анализу подвергались гласные звуки «а», «о», «у» и «и», произносимые в слогах группой из 5 дикторов, в состав которой входили двое мужчин, две женщины и ребенок. Речевой сигнал после оцифровки с тактовой частотой 8000 кГц поступал на обработку. Сигнал после фильтрации различными видами фильтров разделялся на два канала. В одном из каналов производилась его задержка на значения, равные полупериодам частот в диапазоне от 200 Гц до 3500 Гц. После суммирования сигналов они разбивались на участки по 2,5 мс, на которых вычислялся средний уровень сигнала.

На участках длительностью 25 мс с шагом 25 мс фиксировался минимальный уровень, который выводился на результирующий график. В процессе анализа были испытаны варианты использования фильтров верхних частот, полосовых фильтров с полосой пропускания 200 Гц и фильтров, имеющих амплитудно-частотную характеристику (АЧХ), близкую к АЧХ точек основной мембраны. Сравнительные испытания показали, что применение имитаторов фильтров мембраны дает лучший результат.

На рисунке 1 показан амплитудный спектр звука «а» из слога «ао», выполненный полосовыми фильтрами с полосой пропускания 200 Гц на участке длительностью 75 мс. На спектре гармоники импульсов основного тона фактически замаскировали первую форманту, создав неопределенность в распознавании звука.

На рисунке 2 показаны графики функций суммарных сигналов, прошедших мембранные фильтры со средними частотами 750 Гц (форманта F1) и 1275 Гц (форманта F2). Следует отметить, что функция выделила только две фонологические форманты 725 Гц и 1000 Гц, исключив из рассмотрения ложные форманты в области 300 Гц и 2300 Гц, которые идентифицируют звук «и», а также колебания на 450 Гц, 525 Гц и 640 Гц. Так как в амплитудном спектре присутствуют частоты, комбинация которых может принадлежать звукам «у», «о», «а» и «и», автоматизированной системе распознавания пришлось бы проводить дополнительный анализ. Аналогичный результат показала обработка остальных анализируемых звуков различных дикторов.

В процессе выполнения работы было выдвинуто предположение, что указанной обработке подвергаются и согласные звуки, в том числе взрывные, которые опознаются с достаточной степенью надежности по амплитудно-частотному спектру [3]. Для проверки данного предположения был проведен анализ нескольких согласных звуков («С», «Ш», «Д», «Б», «Х») в непрерывной речи.

На рисунке 3 показан спектр области первой форманты звука «Ш» из слова «САША», выполненный полосовыми фильтрами с полосой

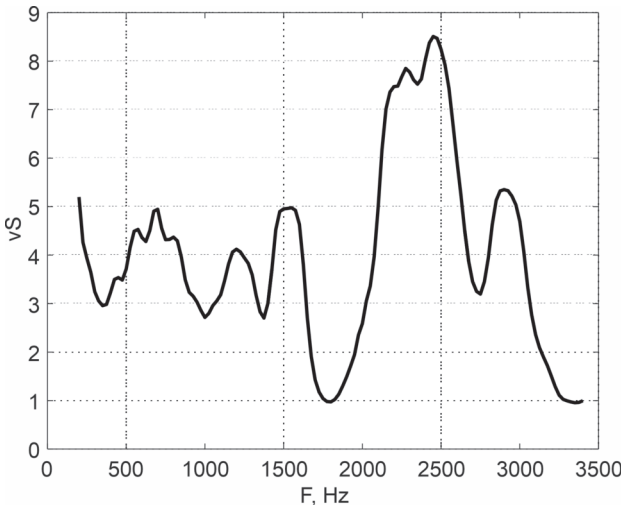

Рисунок 3. Спектр области первой форманты звука «ш» из слова «САША»

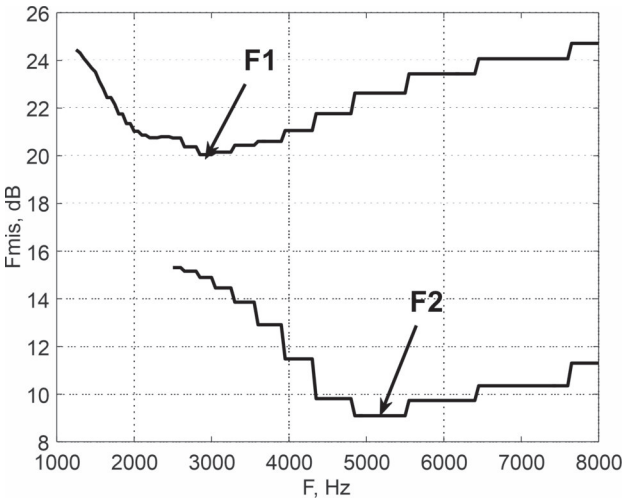

Рисунок 4. График функций суммарных сигналов первой и второй формант звука «ш» из слова «САША»

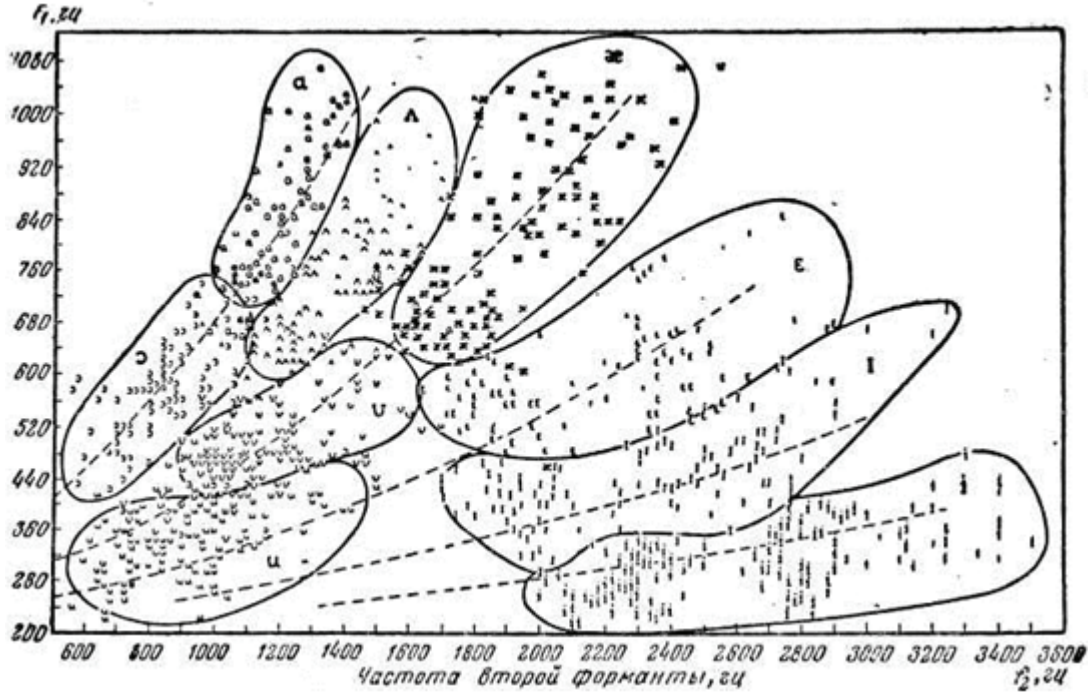

Рисунок 5. Контуры, охватывающие большинство точек зависимости частоты первой форманты от частоты второй для девяти английских гласных звуков

пропускания 200 Гц на участке, длительностью 75 мс. По данному спектру достаточно сложно выделить фонологическую форманту.

На рисунке 4 показаны графики функций суммарных сигналов, прошедших фильтры высоких частот с частотой среза 1500 Гц (форманта F1) и 4100 Гц (форманта F2). Созданные функции выделили только два минимума, соответствующие фонологическим формантам с частотами 2900 Гц и 5200 Гц, что не противоречит положению, что наибольшая степень управляемости акустических характеристик речевого сигнала при сосредоточенном возмущении может быть достигнута лишь относительно пары резонансов [3]. В связи с этим можно предположить, что произведено выделение двух фонологических формант и подобная обработка речевого сигнала может выполняться в слуховом органе уже на первичном этапе обработки.

Приняв за основу предположение, что в речевой информации фонемы кодируются двумя фонологическими формантами и их выделение выполняется в кортиевом органе, рассмотрим возможную связь фонемных областей с физическими характеристиками органов слуха.

При рассмотрении спектрограмм русских звуков от различных дикторов выяснилось, что отклонения от средних спектрограмм подчиняются

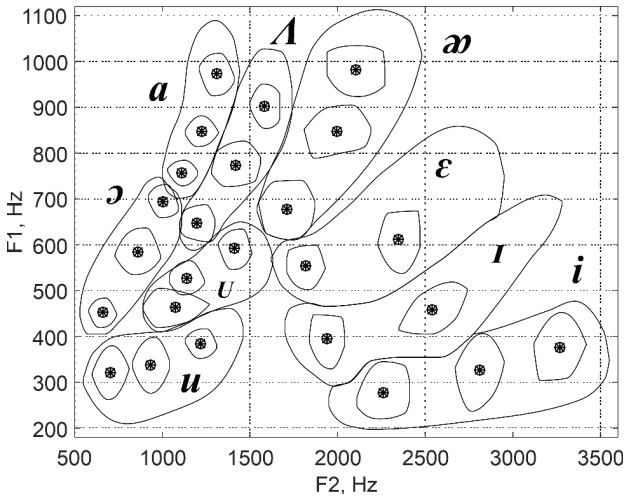

Рисунок 6. Стилизованное представление графика девяти английских гласных звуков

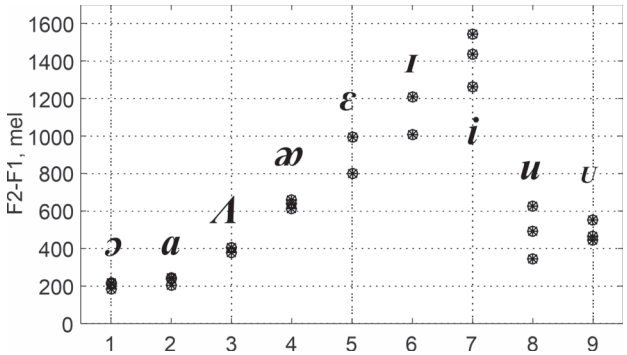

Рисунок 7. График разностей высот тона формант девяти английских гласных звуков

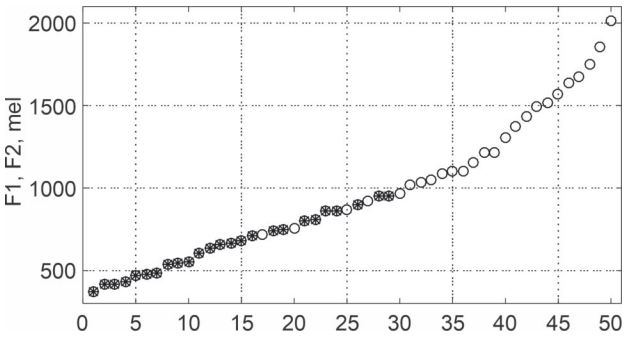

Рисунок 8. Высоты тона первых и вторых формант центров областей девяти английских гласных звуков

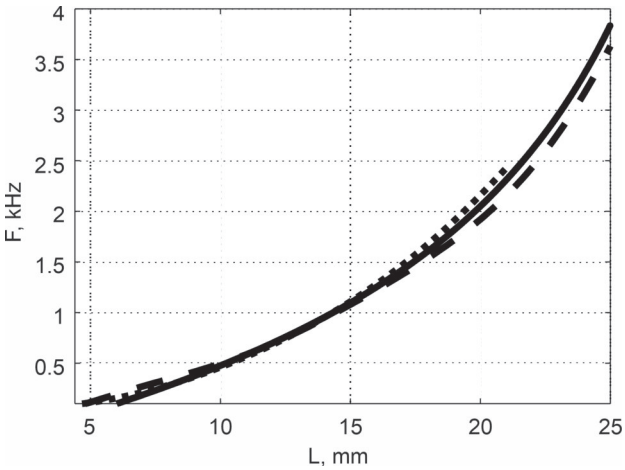

Рисунок 9. Графики расположения частот на мембране кортиевого органа

нормальному закону распределения в каждой из полос, равных в масштабе мел. Таким образом, решение задачи распознавания по существу трактуется как решение задачи выбора одного из полезных сигналов на основе смеси сигнала и помехи [2].

В [9] отмечены контуры областей для девяти английских гласных. Копия указанного графика приведена на рисунке 5. При анализе данного графика можно обратить внимание на тот факт, что координаты пары формант для каждого звука группируются в некоторые области.

Если отметить с некоторой степенью достоверности данные области и их центры, то можем прийти к следующему графику, представленному на рисунке 6.

При рассмотрении рисунка 6 автором было сделано предположение, что одним из факторов распознавания речевых звуков является разность по частоте между первой и второй формантами. С этой целью был составлен график зависимости звука от расстояния между формантами в размерности высоты тона в мелах. Перевод частоты в высоту тона выполнялся по зависимости [2, см. рисунок 5.1]. Результат приведен на рисунке 7.

Из графика видно, что разности формант большинства гласных звуков укладываются в сетку 200 мел. При учете факта достаточно грубого выделения областей звуков и их центров данная зависимость заставляет обратить на себя особое внимание.

На рисунке 8 первые и вторые форманты центров областей размещены по возрастанию их мелодической высоты тона. На данном графике точки, соответствующие первым формантам, зачернены. Из графика видно, что частоты центров в основном имеют различное значение, при этом точки, соответствующие вторым формантам, располагаются в промежутках точек первых формант, соблюдая имеющуюся дистанцию.

Учитывая, что для характеристики звуков речи используются различные системы параметров, оценивающие речь по различным критериям, автор посчитал, что в качестве основы анализа следует использовать геометрические характеристики органов слуха, в частности расположение на кортиевом органе чувствительных клеток.

В литературе опубликовано несколько различных графиков расположения частот на мембране кортиевого органа. Автором было проведено сравнение зависимостей, выведенных на основе измерений амплитудных характеристик смещений базилярной мембраны, выполненных Bekesy G. [4, см. рисунок 4.6], графика пороговых значений девиации частоты [10, см. рисунок. 12.4] и графика естественных шкал основной мембраны внутреннего уха [10, см. рисунок 13.2]. Выведенные зависимости изображены на рисунке 9. На нем сплошной линией изображена зависимость, определенная по графику пороговых значений, штриховой линией – по графику естественных шкал, пунктирной линией – по измерениям Bekesy G. На рисунке 9 график по [10, см. рисунок 12.4] смещен по оси расстояния на 5,5 мм, а график по [10, см. рисунок 13.2] смещен на 3,4 мм.

По графикам видно, что в рассматриваемом диапазоне частот 200…4000 Гц результаты равны с точностью до постоянной. Этот факт может свидетельствовать о том, что используются различные точки отсчета и/или авторы используют различные критерии для определения характерных точек. Учитывая идентичность характеристик, для дальнейшей работы использую функцию, построенную на основе пороговых значений девиации частоты [10].

Перевод графика разностей высот тона формант по рисунку 7 в размерность расстояний показывает, что расстояния между формантами одного звука укладываются в основном в сетку 2 мм. Расстояние в 2 мм, превышающее отрезок влияния, равный 1,3 мм [10], обеспечивает четкое независимое восприятие формант и допускает ошибку в произношении или выделения форманты практически до 100 Гц при правильной идентификации звука речи слуховым органом.

Минимальная величина ступени ощущения высоты тона, определяемая слушателем при прослушивании тестового сигнала, зависит от уровня тестового сигнала. Установлено, что для сигнала с уровнем 80 фон ступень соответствует 37 мкм по основной мембране кортиевого органа [10], а для уровня 70 дБ соответствует 52 мкм [8]. Предполагая, что шаг расположения формант связан со ступенями ощущения высоты тона, на основании графика по рисунку 8, переведенного в расстояния, вычисляю гистограмму расстояний между первыми 37 точками. В связи с ограниченным материалом гистограммы строю для трех параметров: 37 мкм для максимальной чувствительности, 52 мкм в соответствии с [8] и 74 мкм, исходя из предположения, что расстояние между формантами располагается в сетке с шагом, равным двум ощущениям высоты тона по [10]. Полученные гистограммы приведены на рисунке 10. На рисунке гистограмма с шагом 74 мкм изображена сплошной линией, с шагом 52 мкм – точечной, с шагом 37 мкм – штриховой линией.

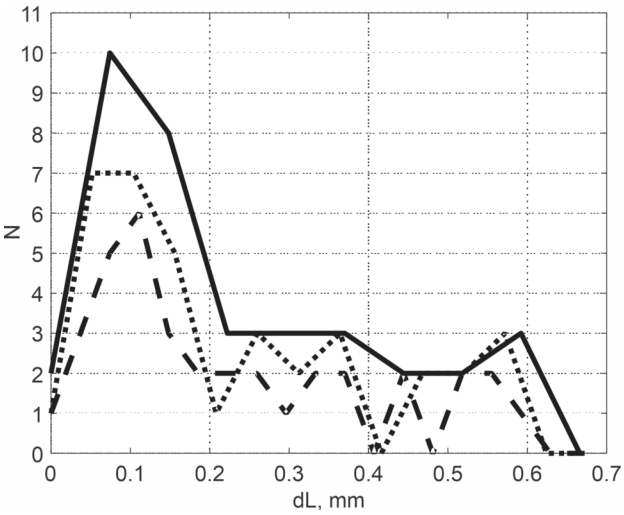

Рисунок 10. Гистограммы расстояний между точками формант для интервалов анализа 37, 52 и 74 мкм

Гистограммы показывают, что наиболее вероятное расстояние между точками базовых формант находится в области 74…104 мкм.

На основе [10] примем, что шагом ощущения высоты тона является фиксированная группа из четырех клеток, занимающая на основной мембране отрезок в 9 × 4 = 36 мкм. Для надежного распознавания высоты тона необходимо, чтобы при слабом звуке расстояние между точками анализа было не меньше удвоенной ступени распознавания высоты тона при понижении чувствительности. Шаг анализа высоты тона принят равным 36 × 3 = 108 мкм достаточно условно по следующим соображениям:

– он в два раза больше 52 мкм, определенных в [8], обеспечивает разделение формант по высоте тона;

– 20 последовательных участков составляют 2,16 мм, которые были приняты ранее за сетку расстояний между формантами одного звука;

– фиксированные группы из четырех клеток с шагом три группы образуют базовую сетку анализа основных формант, удовлетворяющую условиям предыдущих положений.

Исходя из указанных положений, сформирована таблица расположения базовых формант.

При построении таблицы выяснилось, что в сетку 2,16 мм попадают разности формант только 6 английских гласных из 9, но остальные укладываются в сетку 1,08 мм при минимальном расстоянии не менее 3 мм. Дополнительно в таблицу

Таблица. Взаимное расположение базовых формант на основной мембране кортиевого органа

	1	2	3	4	5	6	7	8	9	10
0	И	Б	i1		У		Ы	u1	u2	i2
10			i3	I1	u3		Э		ɔ1	
20	I2	U1		U2			О		ɛ1	
30	U3	ɔ2	I3	ɛ2	У	Λ1	ɔ3	u1	ɔ1	œ1
40	А				ɛ3	a1	О	Λ2		
50	œ2	ɔ2	a2			Λ3	ɔ3		u2	œ3
60	a3	U1				a1				
70	А	Б	a2	U2	u3	Λ1	Х			
80	a3						Ы	Λ2	Я	С
90	U3					Λ3			Д	œ1
100							Э		ɛ1	
110	œ2			I1					Я	œ3
120	И		i1	ɛ2						
130	I2			Ш					Д	i2
140			I3		ɛ3					
150			i3							
160										
170				Ш						С
180										
190							Х			

включены форманты шести гласных звуков русского языка [11] и согласные звуки, параметры которых определены в процессе проведения работ.

С целью лучшей визуализации таблица приведена в размерности принятого шага анализа высоты тона, равного 0,108 мм, и со смещенным началом отсчета расстояния. Цифры рядом с фонетическим знаком обозначают порядковый номер аллофона, подчеркнутые знаки отмечают размещение вторых формант данного аллофона. По частотным координатам гласных звуков в таблице сформирован показанный на рисунке 11 суммарный график с исходными и табличными частотами базовых формант.

На рисунке 11 знаками «+» отображены исходные данные, знаками «*» – русские гласные звуки, знаками «о» – результаты преобразований. Сводный график показывает достаточно высокую степень совпадения частот. Для первой форманты несовпадение не превышает 10 Гц, для второй – не более 80 Гц. Координаты звука ɔ3 были смещены преднамеренно после детального рассмотрения рисунка 5. При выполнении дальнейших работ полученные в результате преобразований фонемы условно назовем базовыми фонемами.

С целью проведения дальнейшего анализа строю график расположения первых формант и расстояний между формантами в размерности принятого шага анализа высоты тона. График представлен на рисунке 12. Анализ данного графика показывает следующее:

– при близком расположении первых формант разность вторых формант разных звуков значительно отличается;

– первые форманты одинакового звука располагаются на некотором расстоянии друг от друга при одинаковой разности формант.

Указанные свойства базовых фонем обеспечивают высокую помехоустойчивость речевого сигнала и правильное выделение звука при значительном различии размеров органов артикуляции у мужчин, женщин и детей. Установлено, что на фонемном уровне потенциально возможно обнаружить около 75 % и исправить около 37,2 % одиночных ошибок и восстановить до 75 % пропущенных артикуляционных признаков [3].

С целью определения возможности распознавания звуков в слитной речи было проведено выделение фонологических формант предложенным методом и вычисление дистанции текущих фонем от базовых фонем различных звуков. График дистанций звуков от базовых фонем в слове «САША» представлен на рисунке 13.

На рисунке сплошной толстой линией изображена текущая дистанция звука «С», штрихпунктирной – звука «А», пунктирной – звука «Ш». Для сравнения тонкой линией изображена текущая дистанция звука «О», отсутствующая в данном слове. С целью наглядности на графике не показаны дистанции произнесенных звуков от базовых фонем остальных звуков. Показания сняты с шагом 25 мс, длительность окна анализа составляет 25 мс.

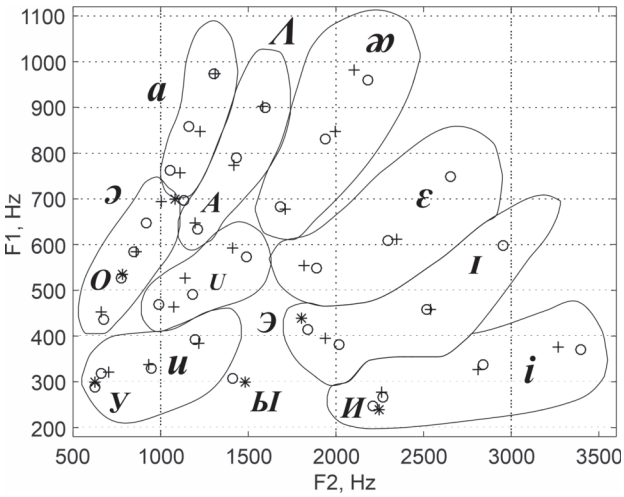

Рисунок 11. Сводный график расположения гласных звуков в координатах частот первой и второй формант

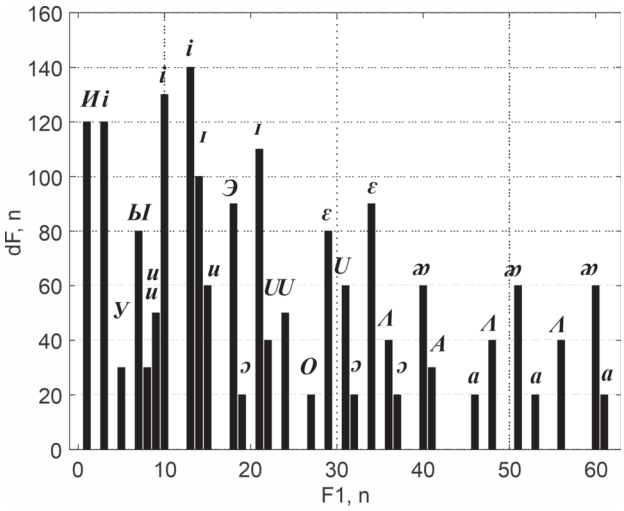

Рисунок 12. График расположения первых формант звуков и расстояний между формантами данного звука

Из графика видно четкое выделение звука по минимальной дистанции, относительная стабильность дистанции на всем участке произношения звука и короткий переход от звука к звуку продолжительностью не более 25 мс.

Автором был проведен ряд экспериментов по анализу йотированных звуков «Я» и «Ё». Спектральный анализ ряда слов с указанными звуками, выполненный полосовыми фильтрами с полосой пропускания 200 Гц, показал наличие изменяемой в течение 100…150 мсек форманты. Для звука «Я» форманта изменялась от 1900 Гц до 1500 Гц с последующей фиксацией в данной области. При этом в течение всего времени произношения звука «Я» на участках длительностью 12,5–25 мс слышался только этот звук. Данный эффект проявлялся как для начального расположения звука (слово «ЯБЛОКО»), так и внутри слова (слово «ДЯДЯ»).

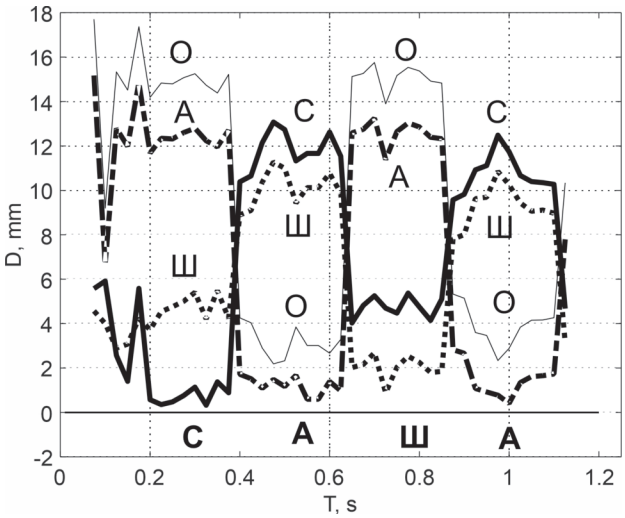

Рисунок 13. График дистанций звуков от базовых фонем в слове «САША»

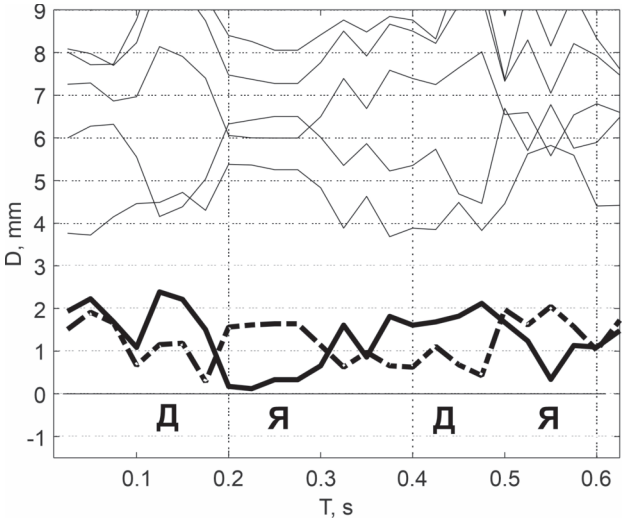

Рисунок 14. График дистанций звуков от базовых фонем в слове «ДЯДЯ»

График дистанций звуков от базовых формант для слова «ДЯДЯ» представлен на рисунке 14. Данный график показывает выделение звука «Я» на всем протяжении его произношения.

Обсуждение результатов

Ряд исследований, выполненных отечественными и зарубежными физиологами [3], констатируют различия в обработке слуховым анализатором человека речевых и неречевых звуков. Следовательно, в речевом сигнале имеется параметр, который отсутствует в естественных природных звуках, в том числе издаваемых живыми организмами, и по которому слуховой анализатор определяет необходимость особой обработки звукового сигнала.

Установлено, что в речевом тракте наибольшая степень управляемости акустических характеристик речевого сигнала может быть достиг-

нута лишь относительно пары резонансов. Если предположить, что эти два резонанса создают пару фонологических формант и их образование и поддержание в требуемом состоянии является целью системы управления по созданию акустического образа фонемы, то получим передачу речевой информации двухтональным многочастотным аналоговым сигналом, аналогичным сигналу DTMF, используемому для набора телефонного номера.

Для безошибочного приема информации необходимо, чтобы кодовое расстояние между ближайшими фонемами было не менее двух частотных групп, равных 90 мелам и соответствующих зонам влияния на кортиевом органе [10]. При соблюдении указанного условия в речевом диапазоне частот может разместиться до 70 взаимно независимых двухчастотных фонем (без учета возможной реализации в речевом аппарате). Количество речевых звуков в большинстве языков находится в диапазоне 35…46 фонем, что может служить одним из подтверждений правильности выбранного направления исследований.

Выявленные достаточно узкие диапазоны частот для первых формант гласных звуков могут служить основанием считать, что для кодирования используется фиксированная сетка частот. На основании изложенного были определены четыре условия, которым должна удовлетворять кодовая таблица речевой информации.

1. Физическая реализуемость при статическом и динамическом состоянии речевого аппарата.

2. Расстояние между двумя частотами одной фонемы не менее двух частотных групп.

3. Кодовое расстояние между ближайшими фонемами не менее двух частотных групп.

4. Сетка частот должна обеспечивать достоверную различимость ближайших частот, то есть расстояние между ними должно быть не менее двух ступеней ощущения высоты тона.

По результатам анализа вариантов размещения кодовой таблицы и с учетом долговременной стабильности ее для субъекта, а также идентичности для разных языков была принята гипотеза о том, что указанная таблица присуща только человеку – при этом каждой частоте соответствует отдельная группа чувствительных клеток, расположенных на определенном расстоянии от геликотремы. С учетом указанных условий и была разработана таблица «Взаимное расположение базовых формант на основной мембране кортиевого органа». При составлении кодовой таблицы использованы материалы [9; 11] и согласные звуки, параметры которых определены в процессе проведения работ.

Звуки в русской речи имеют длительность от 30 до 200 мс [2]. Следовательно, речевая информация передается двухчастотными посылками со стабильными в течение 30–200 мс частотами на фоне общего сигнала с относительно широким спектром частот. Такая комбинация параметров в природных звуках встречается достаточно редко и может служить отличительным признаком речевого сигнала.

Наличие в улитке двух бегущих волн вызывает повышение и уменьшение уровня сигнала в точках мембраны в соответствии с разностью фаз пришедших в данную точку сигналов. При этом для широкополосного сигнала уменьшение или увеличение среднего уровня возникает в точках, соответствующих присутствующим стабильным по частоте сигналам. Указанный способ позволяет выделять частотные посылки различной длительности. В [4; 10] экспериментально установлено, что слух способен не только выделять из спектра частот шума единственную частотную группу, но и определять наличие и местоположение провала в спектре шума. Моделирование данного процесса на слогах и словах русской речи показало:

– наличие для большинства звуков двух сигналов с мало изменяемыми в течение 25–200 мс частотами;

– возможность выделения указанных сигналов методом суммирования и вычитания двух потоков речевого сигнала с различным сдвигом по времени;

– соответствие выделенных частот фонологическим формантам;

– устранение в большинстве случаев коартикуляции и четкое выделение границы фонем при идентификации выделенных фонем по вычисленным дистанциям от базовых фонем;

– стабильную идентификацию фонем, выделяемых при испытаниях йотированных и согласных звуков, базовыми фонемами данного звука.

Согласно «квантовой гипотезе» Стивенса К.Н., каждый класс звуков любого языка порождается множеством конфигураций речевого тракта, относительно которых акустические характеристики устойчивы, то есть мало изменяются при изменении конфигурации тракта в пределах заданного множества форм [3].

Целью системы управления речевым трактом является создание двух фонологических формант требуемой длительности, определяющих заданную фонему. В модели идеальных целей, предложенной Хенке (Henke) в 1966 г. [3], значения признаков задаются скачком и сохраняются в течение некоторого промежутка времени, тогда

как двигательный аппарат непрерывно отрабатывает заданные цели. Случаи коартикуляции, проявляющиеся в акустических характеристиках речевых сигналов, свидетельствуют об ограниченных способностях системы управления компенсировать взаимные возмущения акустических характеристик звуков в слитном потоке речи или об отсутствия потребности в такой компенсации в некоторых случаях.

Однако в любом случае система управления артикуляцией стремится создать такую форму речевого тракта, которая обеспечила бы достижение желаемых акустических характеристик. Возникающие при этом ложные форманты оказываются подвергнутыми модуляциям различного вида и при обработке речевого сигнала не воспринимаются слуховым аппаратом в качестве фонологических формант.

Управляемая коартикуляция, имеющая место при переходах от согласного звука к гласному и при произношении йотированных звуков (в русской речи), обеспечивает формирование для ряда звуков фонологических формант, которые не могут быть получены при статическом состоянии гортани.

Широкополосное возбуждение и многорезонансный голосовой тракт создают широкополосный речевой сигнал, содержащий ложные форманты и переходные процессы, свойственные конкретному диктору. С точки зрения распознавания речи указанные характеристики являются помехой, которую существующие системы распознавания устраняют при помощи сравнения с базой образов помех, содержащих речевую информацию.

По литературным данным, количество сведений, вводимых в начало речеобразующего тракта и управляющих изменениями его конфигурации, не превосходит 50 бит/сек, а для ввода звуковой информации в системы распознавания принято использовать канал со скоростью не менее 64 000 бит/сек. В результате системе приходится выполнять обработку информации, превосходящую полезную почти в 100 раз. Учитывая выявленные различия раздельной обработки речевых и неречевых звуков слуховым анализатором, следует признать, что анализ амплитудного спектра менее всего подходит для распознавания речевого сигнала, о чем свидетельствуют многочисленные трудности в организации данного процесса и отсутствие значимых результатов [1].

Выводы

Результаты представленной работы позволяют сделать следующие выводы.

1. Основой передачи речевой информации в русской речи является двухтональный многочастотный сигнал с фиксированной базовой сеткой частот. Две частоты данного сигнала являются фонологическими формантами фонемы конкретного звука.

2. Диктор при произношении звука настраивает свой речевой аппарат на формирование двух частот, максимально соответствующих требуемым базовым формантам.

3. Слуховой аппарат присваивает принятой фонеме значение базовой фонемы, находящейся от нее на минимальном кодовом расстоянии.

4. Фонологические форманты базовых фонем закреплены за определенными областями кортиевого органа, и их взаимное расположение связано с величинами ступени ощущения высоты тона и отрезками влияния, при этом определенной области соответствует только одна фонологическая форманта.

5. Широкий спектр речевого сигнала несет, кроме фонологических формант, информацию об эмоциях диктора и приспособлении его речевого аппарата к передаче информации с минимально возможными для данного диктора искажениями.

6. Голосовое, шумовое и импульсное возбуждение голосового тракта и управляемая коартикуляция предназначены для формирования максимально возможного количества фонем в ограниченном частотном диапазоне речевого сигнала.

7. Форманты йотированных звуков (русской речи) и ряда согласных звуков не могут быть получены при статическом состоянии гортани и формируются при коартикуляции.

Составленная в процессе выполнения работы таблица, естественно, должна подвергаться уточнениям и дополнениям на основе объективных и полноразмерных испытаний. Заданный в таблице шаг достаточно условен, но принцип привязки фонологических формант к геометрическим характеристикам кортиевого органа и наличие определенного расстояния между базовыми фонемами, связанного с величиной ступени ощущения высоты тона и отрезками влияния, являются очевидными.

Предполагается, что фактическое расположение фонем при соблюдении граничных условий может различаться в зависимости от языка, диалекта и т. п. Вероятно, что основа закладывается до рождения ребенка на основе разговоров матери и формируется примерно до трехлетнего возраста. Для взрослого человека дополнительное формирование базовых фонем происходит при освоении им иностранных языков.

Принимая во внимание, что основную часть спектра речевого сигнала составляют ложные форманты, переходные процессы и индивидуальные особенности дикторов, и учитывая выявленные различия раздельной обработки речевых и неречевых звуков слуховым анализатором [3], автор считает, что анализ амплитудного спектра менее всего подходит для распознавания речевого сигнала. Исследования в данном направлении следует продолжить для выявления реального механизма выделения речи в слуховом аппарате человека и создания соответствующей автоматической системы распознавания.

Литература

1. Кипяткова И.С., Ронжин А.Л. Карпов А.А. Автоматическая обработка разговорной русской речи. СПб.: Санкт-Петербургский институт информатики и автоматизации РАН, 2013. 314 с.
2. Сапожков М.А. Речевой сигнал в кибернетике и связи. М.: Госиздат. по вопросам связи и радио, 1963. 452 с.
3. Сорокин В.Н. Теория речеобразования. М.: Радио и связь, 1985. 312 с.
4. Фланаган Д.Л. Анализ, синтез и восприятие речи. М.: Связь, 1968. 396 с.
5. Тампель И.Б., Карпов А.А. Автоматическое распознавание речи. СПб.: Университет ИТМО, 2016. 140 с.
6. Вокодерная телефония. Методы и проблемы / под ред. А.А. Пирогова. М.: Связь, 1974. 536 с.
7. Галунов В.И., Гарбук В.И. Акустическая теория речеобразования и система фонетических признаков. СПб.: Санкт-Петербургский государственный университет; Санкт-Петербургский НИИ уха, горла, носа и речи. URL: https://studylib.ru/doc/3831855/akusticheskaya-teoriya-recheobrazovaniya-i-sistema (дата обращения: 14.02.2021).
8. Цвикер Э., Фельдкеллер Р. Ухо как приемник информации М.: Связь, 1971. 255 с.
9. Foulkes J.D. Computer identification of vowel types // JASA. 1961. Vol. 33, no. 1. P. 7–11 (приводится по [2]).
10. Фельдкеллер Р., Цвикер Э. Ухо как приемник информации. М.: Связь, 1965. 104 с.
11. Fant G. Acoustic theory of speech production: with calculations based on x-ray studies of Russian articulations // Mouton & Co, s'-Gravenhage, 1960 (приводится по [2]).

Получено 29.03.2021

Лелейтнер Валерий Олегович, зам. главного конструктора АО НПП «Полигон», 450015, Российская Федерация, г. Уфа, ул. Карла Маркса, 37, стр. 1. Тел. +7 905 356-75-38. E-mail: lel@ufacom.ru

EXTRACTION OF PHONEMES FROM THE MERGED SPEECH AND THEIR IDENTIFICATION

Leleytner V.O.
Joint-stock Company «Poligon», Ufa, Russian Federation
E-mail: lel@ufacom.ru

In this paper, we propose a method for determining the location of phonological formants that is not associated with the analysis of the amplitude spectrum and provides their isolation in the merged speech. The formants selected on this basis are grouped into phonemes, the correspondence of which to a particular sound is determined by the minimum distance from the base phonemes. A table of basic phonemes of vowels and several consonants is compiled. The regularity of the location of the basic phonemes and their binding to the coordinates of the location on the main membrane of the cortical organ is revealed. The results of the work are intended for creating autonomous automated speech recognition systems.

Keywords: *continuous speech, speech recognition, phoneme, formant, the organ of Corti*

DOI: 10.18469/ikt.2021.19.2.03

Leleytner Valeriy Olegovich, Joint-stock Company «Poligon», 37, build. 1, K. Marksa Street, Ufa, 450015, Russian Federation; Deputy chief designer. Tel. +7 905 356-75-38. E-mail: lel@ufacom.ru

References

1. Kipjatkova I.S., Ronzhin A.L., Karpov A.A. *Automatic Processing of Spoken Russian Speech*. Saint Petersburg: Sankt-Peterburgskij institut informatiki i avtomatizatsii RAN, 2013, 314 p. (In Russ.)
2. Sapozhkov M.A. *Speech Signal in Cybernetics and Communication*. Moscow: Gosizdat. po voprosam svjazi i radio, 1963, 452 p. (In Russ.)
3. Sorokin V.N. *Theory of Speech Production*. Moscow: Radio i svjaz', 1985, 312 p. (In Russ.)
4. Flanagan D.L. *Analysis, Synthesis and Perception of Speech*. Moscow: Svjaz', 1968, 396 p. (In Russ.)
5. Tampel' I.B., Karpov A.A. *Automatic Speech Recognition*. Saint Petersburg: Universitet ITMO, 2016, 140 p. (In Russ.)
6. Vocoder Telephony. *Methods and Problems*. Ed. by A.A. Pirogova. Moscow: Svjaz', 1974, 536 p. (In Russ.)
7. Galunov V.I., Garbuk V.I. Acoustic theory of speech formation and the system of phonetic signs. Saint Petersburg: Saint Petersburg State University; Saint Petersburg Research Institute of Ear, Throat, Nose and Speech. URL: https://studylib.ru/doc/3831855/akusticheskaya-teoriya-recheobrazovaniya-i-sistema (accessed: 14.02.2021).
8. Tsviker E., Fel'dkeller R. *The Ear as a Receiver of Information*. Moscow: Svjaz', 1971, 255 p. (In Russ.)
9. Foulkes J.D. Computer identification of vowel types. *JASA*, 1961, vol. 33, no. 1, pp. 7–11 (given by [2]).
10. Fel'dkeller R., Tsviker E. *The Ear as a Receiver of Information*. Moscow: Svjaz', 1965, 104 p. (In Russ.)
11. Fant G. Acoustic theory of speech production: with calculations based on x-ray studies of Russian articulations. Mouton & Co, s'-Gravenhage, 1960 (given by [2]).

Received 29.03.2021

ТЕХНОЛОГИИ ТЕЛЕКОММУНИКАЦИЙ

УДК 004.942

МЕТОДЫ И МОДЕЛИ СЕРВИСА РАСПРЕДЕЛЕНИЯ РЕСУРСОВ В КЛАСТЕРАХ С БАЛАНСИРОВКОЙ НАГРУЗКИ ЦЕНТРОВ ОБРАБОТКИ ДАННЫХ

Мочалов В.П., Линец Г.И., Братченко Н.Ю., Палканов И.С.
Северо-Кавказский федеральный университет, Ставрополь, РФ
E-mail: n.b.20062@yandex.ru

Современные центры обработки данных представляют собой комплексные решения по управлению предприятиями и корпорациями, по организации систем обработки и хранения данных, по эффективному распределению программных приложений между доступными ресурсами. Объектом исследования являются кластеры с балансировкой нагрузки центров обработки данных, содержащие определенное множество серверов приложений, файл серверов, систем хранения данных, систему ввода-вывода, связанных между собой системой коммутации и каналами связи. Целью работы является повышение эффективности функционирования виртуализированных центров обработки данных путем разработки методических подходов, рациональных методов и моделей решения задач распределения нагрузки по его аппаратно-программным средствам. Рассматриваются задачи построения взаимосвязанных математических методов и моделей, необходимых для построения специализированного программного обеспечения, способствующего решению задач создания рациональных планов распределения ресурсов центров обработки данных на заданном временном интервале. В основу предложенных методов и моделей положена система корректных отображений совокупности параметров запросов на известные характеристики физических ресурсов центров обработки данных. В качестве критерия оптимизации используются параметры максимизации производительности системы за определенный временной интервал.