УДК 534.32 **Физико-математические науки**

***Горячкин Б. С.**, кандидат технических наук, доцент; Московский государственный технический университет им. Н.Э. Баумана;*
E-mail: bsgor@mail.ru
***Коробко Д. О.**, магистрант; Московский государственный технический университет им. Н.Э. Баумана;*
E-mail: domitorikor@gmail.com

ВЫБОР ОПТИМАЛЬНЫХ ВОКАЛЬНЫХ ХАРАКТЕРИСТИК ГОЛОСОВОГО БАНКА НА ОСНОВЕ ЭРГОНОМИЧЕСКОГО АНАЛИЗА

Аннотация: В данной статье описаны вопросы выявления оптимальных вокальных характеристик голосового банка (ГБ) для программы-синтезатора речи. Разрабатываются рекомендации, которые позволят уменьшить количество рассматриваемых вариантов при выборе ГБ на основе исследования взаимодействия пользователя, системы и рабочей среды. Производится анализ проблемных аспектов предметной области исследования для определения альтернативных вариантов и основных критериев выбора, таких как, общий обзор технологии синтеза речи, характеристики естественного и искусственного речевого сигнала, особенности слухового восприятия речи и влияние окружающего шума на степень восприятия речи. В результате исследования проведен выбор оптимального класса ГБ для программы-синтезатора речи на основе предложенного метода.

Ключевые слова: конкатенативный синтез речи, голосовой банк, различимость формант, модель человеческого голоса, частотное восприятие речи, шумоустойчивость речи.

Annotation: This article describes the issues of identifying the optimal vocal characteristics of a voice bank (VB) for a speech synthesizer program.

Recommendations are being developed that will reduce the number of options considered when choosing a VB based on a study of the interaction of the user, system and work environment. An analysis is made of the problematic aspects of the research subject area to determine alternative options and basic selection criteria, such as a general overview of speech synthesis technology, characteristics of natural and artificial speech signal, features of auditory speech perception and the influence of ambient noise on the degree of speech perception. As a result of the study, the optimal VB class for a speech synthesizer program was selected based on the proposed method.

Key words: concatenative speech synthesis, voice bank, formant distinguishability, human voice model, frequency perception of speech, noise resistance of speech.

Введение

Задача выбора оптимальных вокальных характеристик голосового банка для программы-синтезатора речи решает проблему создания удобных и доступных методов и средств эргономической оценки. Если не применять к этой задаче некоторые допущения, она является чрезвычайно комплексной – количество ГБ и их параметров слишком велико, чтобы решать эту задачу прямыми способами. В связи с этим, во-первых, следует попытаться упростить задачу, и во-вторых, следует выделить некоторые критические параметры, которые должны стать основными критериями выбора ГБ. Наиболее важным параметром ГБ представляется его вокальный диапазон, т.е. диапазон частот, в котором звучит голос. На основании этой информации можно классифицировать каждый ГБ по типу голоса и распространять результаты анализа классов на входящие в них ГБ. Возможно упрощение задачи за счет:

- Сокращения количества рассматриваемых вариантов;
- Использования методов логического вывода.

С другой стороны, для определения основных критериев выбора следует сфокусироваться на проблемах, связанных с восприятием речи, поскольку они

представляют наибольший интерес с точки зрения эргономического анализа. К этим проблемам относятся:

- Разборчивость отдельных речевых единиц в ГБ;
- Особенности восприятия речи при помощи слуха;
- Возрастные изменения слуха;
- Устойчивость речи к негативному воздействию окружающих шумов.

Перечисленные проблемы затрагивают различные аспекты вербальной коммуникации – процесса передачи информации от одного человека к другому посредством речи. Так, первая проблема относится к формированию речевого сигнала, вторая и третья проблемы – к его восприятию, а четвёртая проблема – к его передаче по среде распространения.

Анализ проблемных аспектов предметной области исследования

Программы-синтезаторы речи

В широком смысле, синтезаторы речи – это системы, предназначенные для преобразования печатного текста в устную речь. Под это определение попадают как аппаратные, так и программные системы, однако в настоящее время аппаратных систем синтеза речи практически не осталось. Кроме того, под это определение попадают как автоматизированные, так и полностью автоматические средства преобразования текста в речь. В узком смысле под синтезаторами речи понимаются исключительно *автоматические программные* системы, способные формировать речевой сигнал по *произвольному* печатному тексту.

Входными данными в алгоритме синтеза речи является письменная речь, т.е. *текст*, а выходными данными – устная речь, т.е. *звук*. Между начальным и конечным состоянием существует несколько промежуточных. Всего в процессе синтеза речи выделяют три основных этапа, последовательно связанных между собой таким образом, что выходные данные предыдущего этапа являются входными данными для последующего [3]:

1. Текст в слова;
2. Слова в фонемы;

3. Фонемы в звуки.

Современные системы синтеза речи (ССР) подразделяют на два класса на основе различий в реализации последнего этапа синтеза речи. Выделяют *конкатенативные* и *параметрические* системы синтеза речи. В конкатенативных (или компилятивных) синтезаторах речи для получения звуков фонем используются фрагменты аудиозаписей речи живых людей. Эти фрагменты хранятся в т.н. голосовых банках (ГБ), содержащих речь отдельного диктора. В параметрических (или формантных) синтезаторах речи моделируется естественная артикуляция посредством компьютерной генерации формант – набора гармонических звуковых колебаний различных частот, формирующих звуки речи [2].

Системы конкатенативного типа более распространены в настоящее время по двум причинам: во-первых, для их создания от разработчиков требуется меньше знаний в области фонетики, чем в случае параметрических синтезаторов речи, а во-вторых, поскольку в конкатенативных системах по определению за основу берётся речь настоящего человека, синтезируемая ими речь обычно воспринимается более естественной, что сказывается на субъективной оценке её качества [9; 10].

Естественный речевой сигнал

Под *естественным речевым сигналом* (ЕРС) понимается звук, который образуется с помощью артикуляционного аппарата человека (гортань, язык и прочее) с целью общения. По аналогии с этим определением можно сказать, что звук, создаваемый с помощью синтезаторов речи, называется *искусственным речевым сигналом* (ИРС). Поскольку задачей синтеза речи является имитация человеческого голоса, к ИРС применимы те же понятия, что и к ЕРС. Голос человека и машины имеет одинаковые характеристики: *высота*, *громкость*, *тембр*. Все они отражают параметры, воспринимаемые субъективно. Так, высота звука – это его воспринимаемая частота, громкость – это интенсивность звука, действующего на барабанные перепонки слушателя, а тембр – специфический обертоновый окрас звука, отличающий один голос от другого.

Для разделения голосов по частотному диапазону, в котором они звучат, можно воспользоваться классификацией, представленной в теории пения. Всего выделяется шесть основных типов голосов, по три типа для мужских и женских, от самого низкого голоса до самого высокого: для мужских голосов это *бас*, *баритон* и *тенор*, а для женских – *контральто*, *меццо-сопрано* и *сопрано* [7].

Представленная классификация может быть распространена не только на певческий, но и на обычный, разговорный голос человека – не только певцов, но и далёких от пения людей. По оценкам певцов и музыкальных педагогов, разговорный диапазон при спокойной речи составляет примерно одну десятую часть от певческого диапазона голоса человека, а при выразительной, эмоциональной речи – около одной четверти. Частота основного тона у мужчин при разговоре обычно находится в диапазоне от 85 до 180 Гц, а у женщин – в диапазоне от 165 до 255 Гц.

Подобное разделение голосов по частотному диапазону позволяет производить сравнительный анализ и решать задачи выбора для различных *классов голосов*. Благодаря этому, к примеру, вместо того, чтобы оценивать и сравнивать каждый отдельный голос, достаточно определить его принадлежность к определённому классу, и если параметры выбранного класса удовлетворяют заданным требованиям, то можно сделать вывод, что и все входящие в этот класс голоса также удовлетворяют этим требованиям.

В целом звуки речи подразделяются на шумы и тоны: тоны в речи возникают в результате колебания голосовых складок, а шумы образуются в результате непериодических колебаний выходящей из лёгких струи воздуха. Тонами являются обычно гласные; почти же все глухие согласные относятся к шумам, а звонкие и сонорные согласные образуются путём слияния шумов и тонов. Отдельные шумы и тоны, подобно речи в целом, исследуются по их высоте, тембру, силе и многим другим характеристикам – их изучением занимается фонетика. С тонами связано такое понятие, как форманта, упомянутое ранее при рассмотрении параметрических ССР.

Термин форманта обозначает определенную частотную область, в которой в следствие резонанса усиливается некоторое число гармоник основного тона, производимого голосовыми связками, то есть в спектре звука форманта является достаточно отчетливо выделяющейся областью усиленных частот. Считается, что для характеристики звуков речи достаточно выделения двух первых формант – FI и FII, которые нумеруются в порядке возрастания их частоты. Для разных тоновых звуков речи характерны определённые частотные диапазоны формант. Например, для фонемы «i» средние значения формант FI = 240 Гц, FII = 2400 Гц, а для фонемы «e» – FI = 390 ГЦ, FII = 2300 Гц.

В зависимости от особенностей голоса каждого отдельного человека, значения формант могут смещаться в сторону более низких или высоких частот, и в случае голосов слишком низких (бас) и слишком высоких (сопрано) это смещение может приводить к тому, что отдельные фонемы, имеющие близкие значения формант, начинают звучать (и, соответственно, слышаться) схожим образом. Подобное явление проявляется слабо или не проявляется вовсе в других типах голосов. Наблюдаемая зависимость между *различимостью формант* и типом голоса представлена в табл. 1 в виде вербальной и числовой шкалы [4].

Таблица 1. Зависимость между различимостью формант и типом голоса

Тип голоса	Различимость формант	
	Значение	Балл
Бас	Хорошо	0,8
Баритон	Очень хорошо	1,0
Тенор	Очень хорошо	1,0
Контральто	Очень хорошо	1,0
Меццо-сопрано	Хорошо	0,8
Сопрано	Средне	0,6

Для изучения спектра речевого сигнала применяются различные статистические методы, одним из которых является усреднение по времени. Этот метод позволяет представить спектрограмму сигнала – трёхмерный график зависимости уровня звукового давления (УЗД), выражаемого в децибелах, от

частоты и времени – в виде двухмерного графика зависимости УЗД от частоты на определённом конечном временном интервале. Это позволяет наблюдать зависимости в речевом сигнале, характерные для речи конкретного диктора и слабо обусловленные содержанием речи. При наличии подобных графиков для нескольких различных дикторов, становится возможно также производить их усреднение по УЗД, что позволяет получить *обобщённую модель человеческого голоса*. Однако если предварительно разбить их на категории по признаку принадлежности голоса диктора к тому или иному типу (см. классификацию выше), становится возможным построить *частную модель голоса определённого типа* – баса, тенора, меццо-сопрано и т.д. Модели мужских голосов разных типов представлены на рис. 1, а женских – на рис. 2.

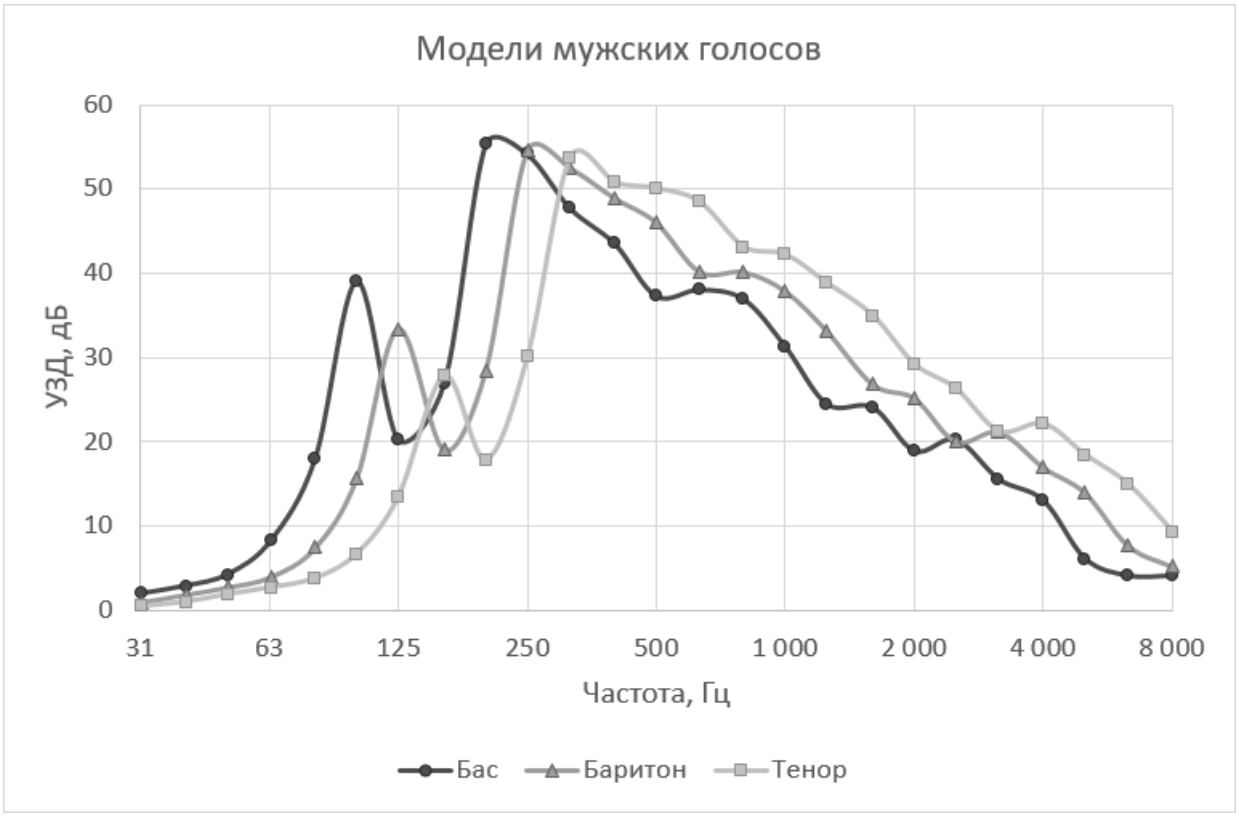

Рисунок 1. Модели мужских голосов

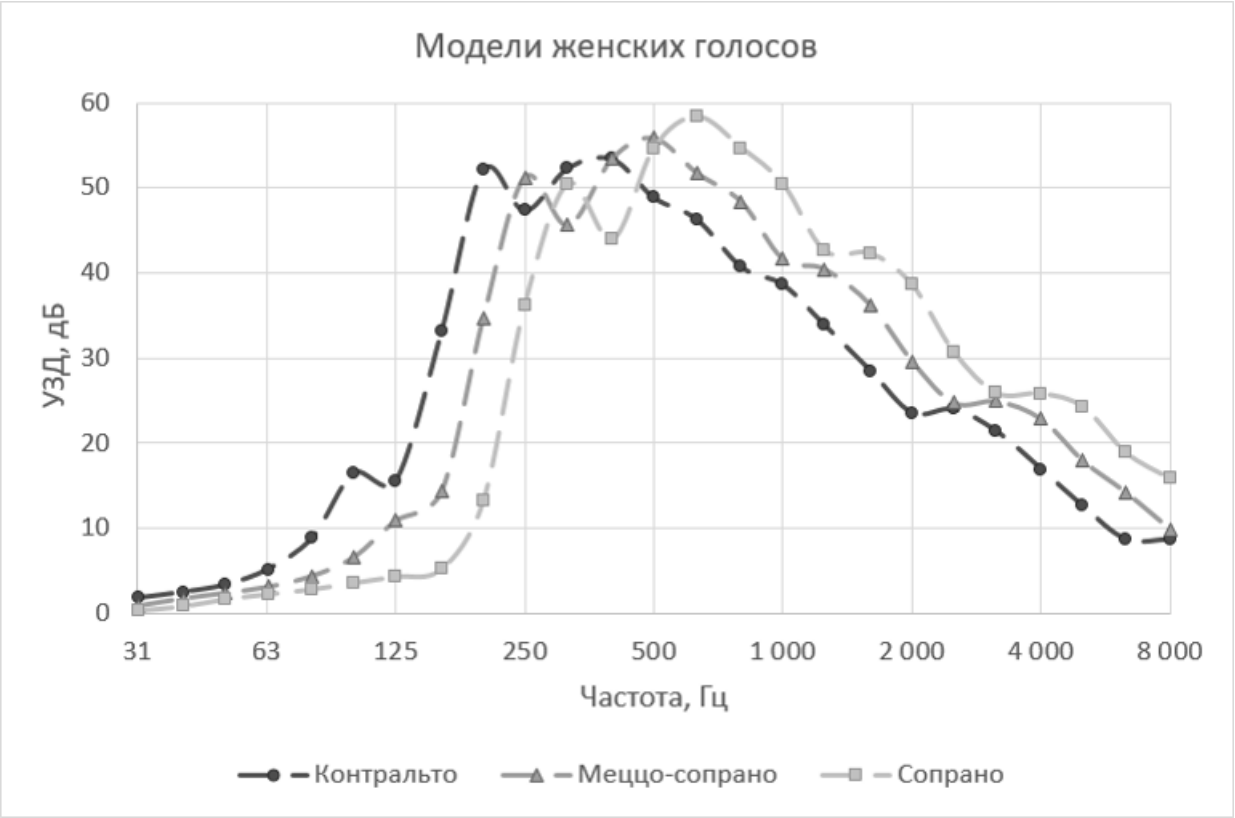

Рисунок 2. Модели женских голосов

Как видно из графиков, первый значимый максимум функции приходится на область частот, в которой расположен основной тон F0, а второй максимум приходится на форманты FI и некоторые форманты FII. Последующие форманты и оставшаяся часть формант FII постепенно теряют звуковую энергию, снижаясь до едва различимого уровня в 10 дБ примерно на частоте 8 кГц.

Особенности восприятия звука

Человек способен слышать звук в пределах от 16 Гц до 20 кГц при передаче колебаний по воздуху и до 220 кГц при передаче звука по костям черепа. Звуковые волны в диапазоне от 100 до 4000 Гц соответствуют человеческому голосу и называются речевыми частотами. Согласно закону Вебера – Фехнера, интенсивность ощущения чего-либо, в том числе и звука, прямо пропорциональна логарифму интенсивности раздражителя. В связи с этим, в оптике и акустике для удобства расчётов, связанных с человеческим восприятием, часто применяется логарифмическая величина *бел* (и, в частности, её дольная единица *децибел*), выражающая отношение двух значений энергетической величины десятичным логарифмом этого отношения.

Использование децибелов при указании громкости звука обусловлено человеческой способностью воспринимать звук в очень большом диапазоне изменений его интенсивности. Диапазон величин звукового давления от минимального порога слышимости звука человеком (20 мкПа) до максимального, вызывающего болевые ощущения, составляет примерно 120 дБ. Например, утверждение «громкость звука составляет 30 дБ» означает, что интенсивность звука в 1000 раз превышает порог слышимости звука человеком.

Помимо интенсивности, громкость главным образом функционально зависит от частоты звуковых колебаний. Для оценки уровня громкости звука была введена логарифмическая единица измерения, называемая *фоном*, которая отличается от шкалы децибелов тем, что в ней значения громкости коррелируются с чувствительностью человеческого слуха на разных частотах. У чистого тона с частотой 1000 Гц уровень в фонах численно равен уровню в децибелах, для других частот используют поправки из таблицы или специального графика – *контура равных громкостей*, представляющего собой семейство кривых, называемых также *изофонами*, и описываемого международным стандартом ISO 226 [6]. Графики стандартных кривых равной громкости чистых тонов приведены на рис. 3.

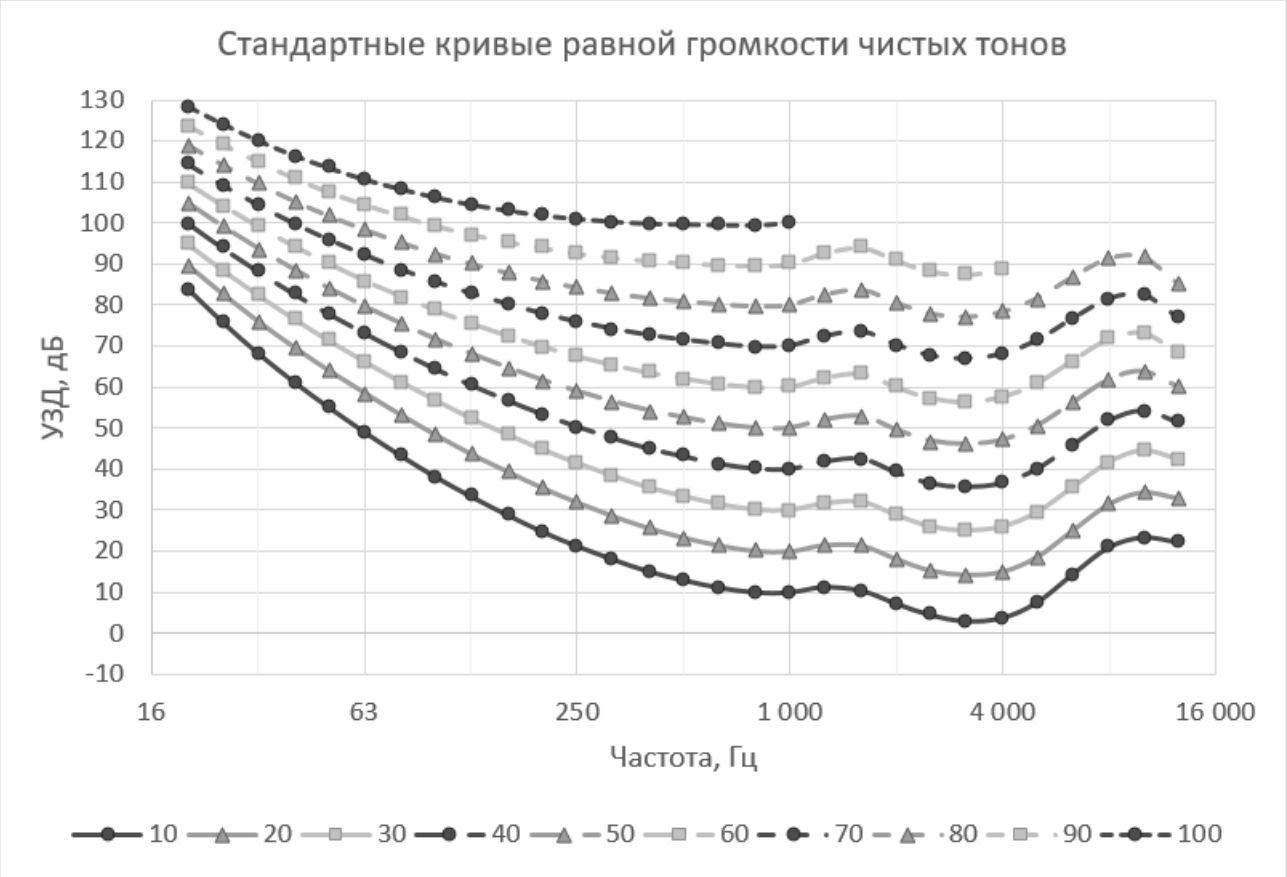

Рисунок 3. Стандартные кривые равной громкости чистых тонов

Как видно из графика, на левой границе речевого диапазона (100 Гц) уровень звукового давления (УЗД) на кривых равной громкости выше, чем на правой границе (4 кГц), что означает, что низкие звуки речи слышатся тише, чем высокие.

Данные, используемые для построения изофон, были получены путём проведения экспериментов с участием множества молодых людей в возрасте до 25 лет, не имеющих нарушения слуха. С возрастом чувствительность человеческого уха к высокочастотным звукам постепенно падает. В возрасте 60 лет в диапазоне речевых частот наибольшие изменения наблюдаются на кривых равной громкости, соответствующие нижним уровням. Возрастные нарушения слуха начинают значительно влиять на восприятие звуков с частотой более 2 кГц, постепенно уменьшая громкость воспринимаемых звуков с увеличением их частоты. Графики смещённых изофон для людей в возрасте 60 лет представлены на рис. 4.

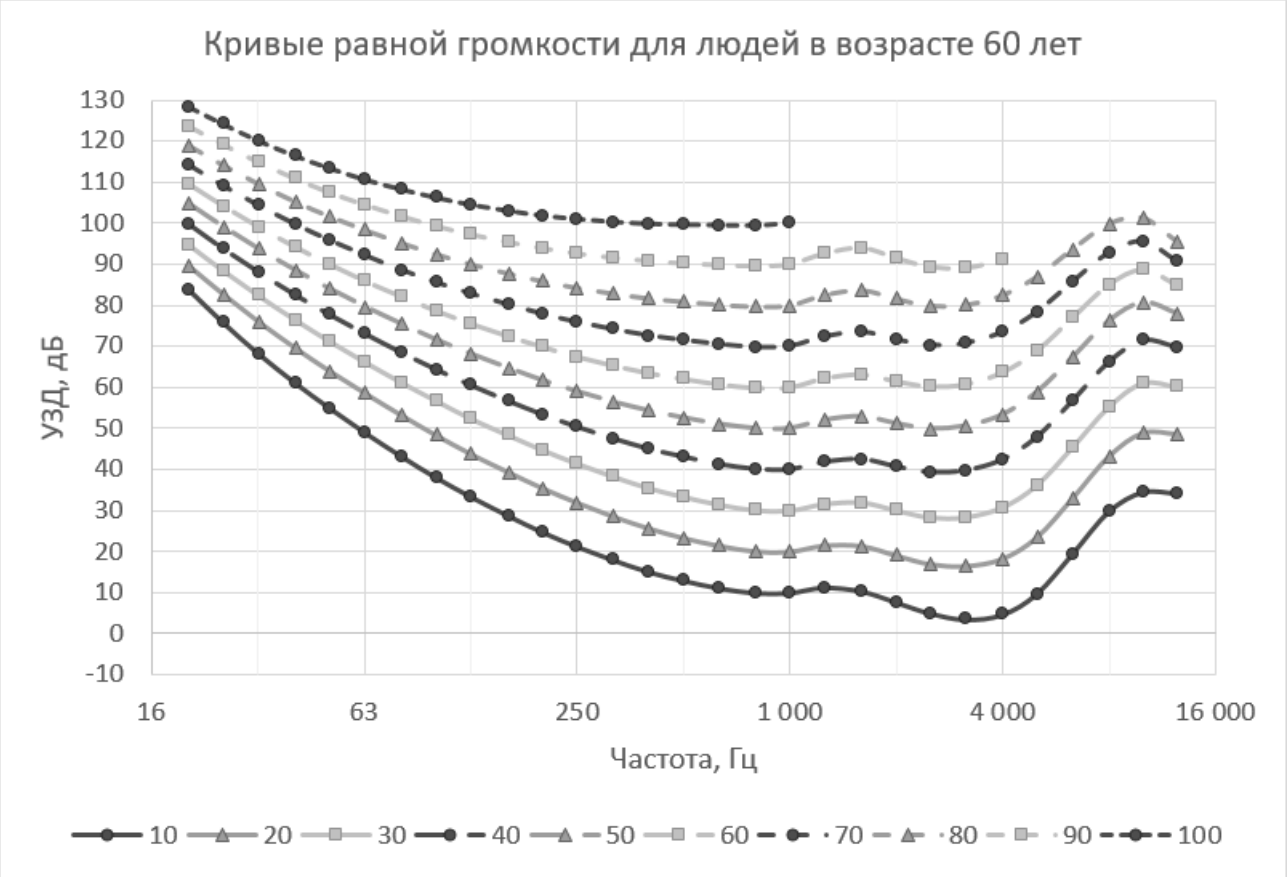

Рисунок 4. Смещённые кривые равной громкости чистых тонов для людей в возрасте 60 лет

Для оценки слухового восприятия речевого сигнала можно воспользоваться *коэффициентом приёма*, выражающим отношение эквивалентного уровня принимаемого сигнала (в данном случае – воспринимаемая речь) к эквивалентному уровню передаваемого сигнала (произносимая речь), рассчитав их значение в речевом диапазоне; он учитывает как частотное распределение сигнала, так и частотное восприятие звуков. Формула для расчёта коэффициента приёма имеет следующий вид:

$$K = L_{\text{экв,прин}} / L_{\text{экв,пер}}, \qquad (1)$$

где эквивалентный уровень звука вычисляется по формуле:

$$L_{\text{экв}} = 10 * \log_{10} \sum_{i=1}^{n} \langle p_{f_i} \rangle_t = 10 * \log_{10} \sum_{i=1}^{n} 10^{\langle p_{f_i,\text{дБ}} \rangle_t / 10}, \qquad (2)$$

где $\langle p_{f_i} \rangle_t$ – усреднённое по времени значение интенсивности звука, измеряемое на частоте f_i, n – количество отсчётов.

Для расчёта $L_{\text{экв,прин}}$ в качестве принимаемого на определённой частоте f_i сигнала можно рассматривать передаваемый речевой сигнал,

скорректированный путём сложения значения его УЗД для данной частоты с соответствующим значением на обратной кривой равной громкости с уровнем 40 фон (рассчитывается как 40 дБ минус значение изофоны). Причина, по которой выбирается именно изофона с уровнем 40, заключается в том, что на неё опирается самая распространённая коррекция для расчёта уровня громкости – коррекция А.

На основе данных, использовавшихся для построения моделей мужских (рис. 1) и женских (рис. 2) голосов, а также стандартных (рис. 3) и смещённых (рис. 4) изофон, можно произвести расчёты $L_{\text{экв,пер}}$, а также $L_{\text{экв,прин}}$ для каждого типа голоса как для случая с *нормальным*, так и для случая с *нарушенным слуховым восприятием речи*, после чего можно рассчитать соответствующие значения коэффициентов приёма $K_{\text{норм}}$ и $K_{\text{нар}}$. Результаты этих расчётов приведены в табл. 2.

Таблица 2. Нарушенное и нормальное слуховое восприятие голосов разных типов

Тип голоса	$L_{\text{экв,пер}}$, дБ	$L_{\text{экв,прин,норм}}$, дБ	$L_{\text{экв,прин,нар}}$, дБ	$K_{\text{норм}}$	$K_{\text{нар}}$
Бас	58,52	48,51	48,49	0,8290	0,8287
Баритон	57,84	51,03	51,01	0,8822	0,8818
Тенор	57,59	53,52	53,50	0,9297	0,9289
Контральто	58,75	53,06	53,05	0,9032	0,9030
Меццо-сопрано	60,14	56,85	56,83	0,9452	0,9450
Сопрано	61,93	60,44	60,42	0,9759	0,9755

Расчёты демонстрируют следующее: эквивалентный уровень громкости воспринимаемой речи в случае с нарушенным слухом $L_{\text{экв,прин,нар}}$ отличается от соответствующего уровня в случае с нормальным слухом $L_{\text{экв,прин,норм}}$ незначительно – различие составляет порядка 0,01-0,02 дБ. Причина этого заключается в том, что эквивалентный уровень громкости – это логарифмический показатель, учитывающий значения уровней громкости звука

в широком диапазоне частот, и поскольку возрастные нарушения слуха проявляются в основном на частотах выше 2 кГц, это слабо отражается на общей форме воспринимаемого речевого сигнала. В связи с этим значения коэффициентов приёма и $K_{\text{нар}}$ и $K_{\text{норм}}$ также отличаются на весьма малые величины. Однако, несмотря на малый масштаб изменения, коэффициент приёма всё же отражает влияние возрастных нарушений слуха на восприятие речи.

Влияние шума на восприятие речи

Окружающий человека шум оказывает влияние на восприятие речи – маскирование шумами речевого сигнала приводит к снижению его разборчивости.

Допустимые уровни шума в окружающей человека обстановке определяются стандартами и нормами. В России для определения допустимого уровня шума на рабочих местах, в жилых помещениях, общественных зданиях и на территории жилой застройки используются ГОСТ 12.1.003-2014. ССБТ «Шум. Общие требования безопасности» и СН 2.2.4/2.1.8.562-96 «Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки» [5; 8]. Нормирование шума в слуховом диапазоне осуществляется по предельному спектру уровня шума. Этот метод устанавливает предельно допустимые уровни звука (ПДУЗ) в девяти октавных полосах со среднегеометрическими значениями частот 31,5, 63, 125, 250, 500, 1000, 2000, 4000 и 8000 Гц. График распределения ПДУЗ для нескольких различных условий среды представлен на рис. 5.

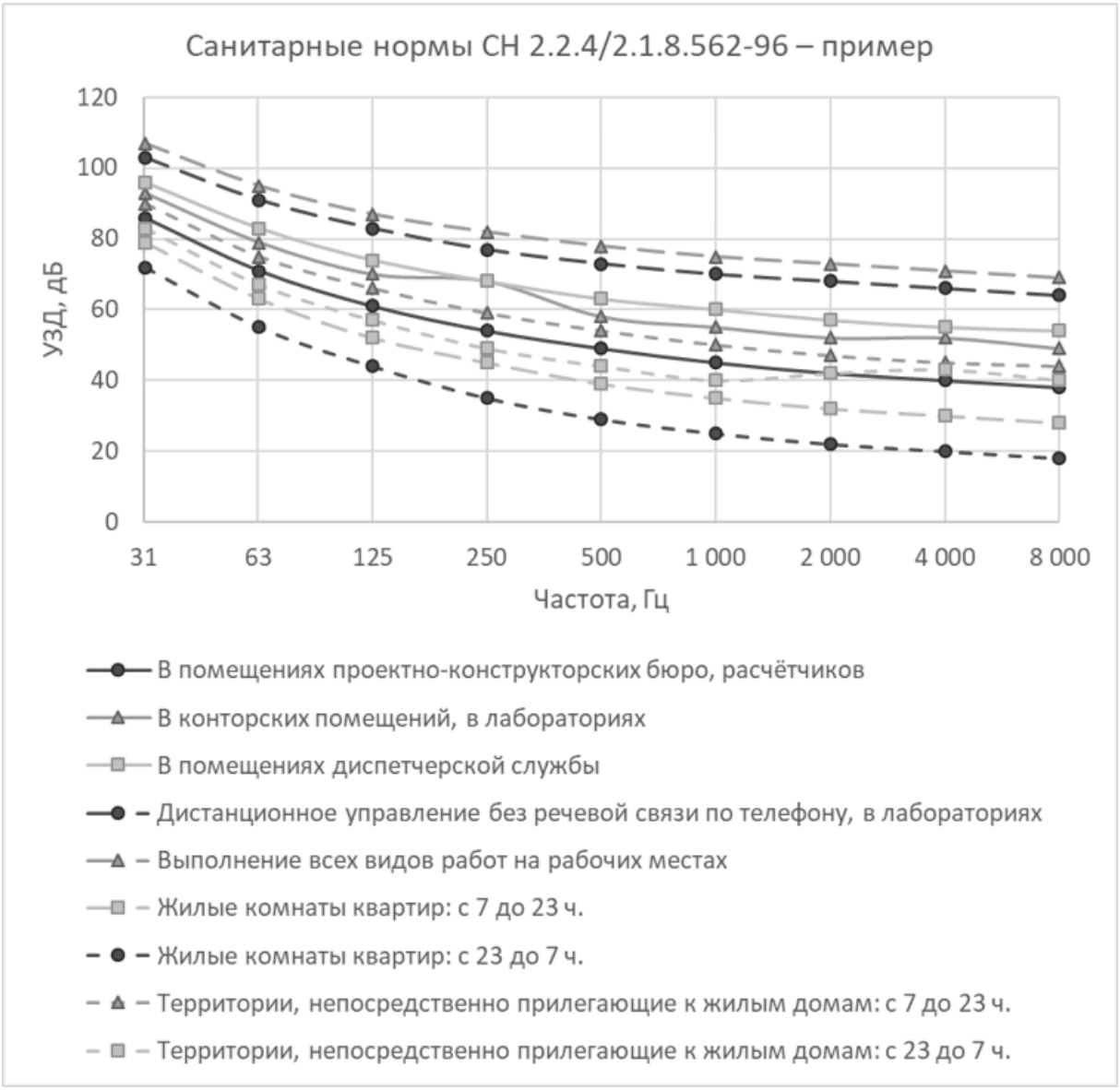

Рисунок 5. Распределение ПДУЗ для различных условий среды

Установленные данными документами ПДУЗ могут рассматриваться как значения максимально неблагоприятные, т.е. при любых иных условиях на каждой рассматриваемой октавной полосе частот в каждой конкретной среде может наблюдаться уровень звука, не превышающий указанное значение. Это допущение позволяет использовать санитарные нормы как оценку сверху для УЗД шумов в заданных условиях среды, и благодаря этому можно использовать ПДУЗ в расчётах, связанных с определением степени влияния шумов на восприятие речи.

Под разборчивостью речи (РР) понимается процентное отношение правильно распознанных единиц речи: слогов, слов или фраз, в зависимости от

метода оценки. Существует несколько способов различных методов оценки РР, применяющихся на практике. Они подразделяются на эмпирические, модуляционные и формантные. Формантные методы являются наиболее объективными, они основаны на разбиении спектра речи на полосы частот (октавных, третьоктавных, равноартикуляционных или иных, произвольной ширины), в пределах каждой из которых плотность вероятностей формант можно считать неизменными, а РР оценивается для разных соотношений сигнал/шум [1].

Отношение сигнал/шум (ОСШ) является безразмерной величиной, часто выражаемой в децибелах, определяемой по формуле:

$$SNR_{\text{дБ}} = 10 * \log_{10}\left(P_{\text{сигнал}}/P_{\text{шум}}\right) = 20 * \log_{10}\left(A_{\text{сигнал}}/A_{\text{шум}}\right) =$$
$$= P_{\text{сигнал,дБ}} - P_{\text{шум,дБ}} = A_{\text{сигнал,дБ}} - A_{\text{шум,дБ}}, \qquad (3)$$

где P – энергетическая величина (интенсивность звука), A – силовая величина (звуковое давление).

В данной работе ОСШ используется в качестве количественной оценки влияния шумов окружающей среды на передаваемый речевой сигнал, соответствующий определённому типу голоса. Рассматриваются девять различных вариантов условий среды, ПДУЗ для которых были представлены на рис. 5. Для каждого типа голоса, распределение УЗД по частотам для которых представлено на рис. 1 и рис. 2, производится расчёт *среднего значения ОСШ для различных условий среды на третьоктавных полосах частот* (со среднегеометрическим значением от 31,5 до 8000 Гц) по следующей формуле:

$$SNR_{k,\text{дБ}} = 1/m * 1/n * \sum_{j=1}^{m}\sum_{i=1}^{n} SNR_{ijk,\text{дБ}}, \qquad (4)$$

где k – тип голоса, m – количество различных условий среды ($m = 9$), j – условия среды, n – количество диапазонов частот ($n = 25$), i – номер диапазона частот, $SNR_{ijk,\text{дБ}}$ – среднее значение ОСШ на диапазоне i при условиях среды j для типа голоса k.

Результаты этих расчётов приведены в табл. 3.

Таблица 3. Устойчивость голосов разных типов к зашумлению

Условия среды	$SNR_{jk,\text{дБ}}$					
	Бас	Барит.	Тенор	Конт.	Мецц.	Сопр.
В помещениях проектно-конструкторских бюро, расчётчиков	-29,55	-29,23	-29,03	-27,60	-27,47	-27,46
В конторских помещений, в лабораториях	-39,63	-39,31	-39,11	-37,68	-37,55	-37,54
В помещениях диспетчерской службы	-43,39	-43,07	-42,87	-41,44	-41,31	-41,30
Дистанционное управление без речевой связи по телефону, в лабораториях	-52,91	-52,59	-52,39	-50,96	-50,83	-50,82
Выполнение всех видов работ на рабочих местах	-57,59	-57,27	-57,07	-55,64	-55,51	-55,50
Жилые комнаты квартир: с 7 до 23 ч.	-20,27	-19,95	-19,75	-18,32	-18,19	-18,18
Жилые комнаты квартир: с 23 до 7 ч.	-10,99	-10,67	-10,47	-9,04	-8,91	-8,90
Территории, непосред-но прилегающие к жилым домам: с 7 до 23 ч.	-34,43	-34,11	-33,91	-32,48	-32,35	-32,34
Территории, непосред-но прилегающие к жилым домам: с 23 до 7 ч.	-27,07	-26,75	-26,55	-25,12	-24,99	-24,98
$SNR_{k,\text{дБ}}$	*-35,09*	*-34,77*	*-34,57*	*-33,14*	*-33,02*	*-33,00*

Как видно из результатов расчётов, для всех типов голосов шумы окружающей среды в наиболее неблагоприятных (предельно допустимых) условиях в среднем оказываются сильнее, чем речевой сигнал, что выражается в отрицательном значении ОСШ. Однако полученные результаты всё же позволяют оценить и соотнести устойчивость разных типов голосов к

зашумлению: чем меньше значение ОСШ, тем сильнее теряется сигнал на фоне шумов.

Определение направления исследования

Анализ предметной области позволяет на основе логического вывода, во-первых, разбить множество ГБ на подмножества-классы по определённому критерию – вокальному диапазону, – а во-вторых, распространить результаты исследования классов на входящие в них банки. Таким образом, в решаемой задаче можно выделить следующие шесть альтернативных вариантов-типов голосов: бас, баритон, тенор, контральто, меццо-сопрано и сопрано.

Рассмотренные в предыдущем разделе аспекты позволили определить основные критерии сравнения. Они представлены в табл. 4.

Таблица 4. Критерии сравнения

Обозначение	Наименование	Характеристика
К1	Различимость формант	Экспертная оценка
К2	Нормальное слуховое восприятие речи	Коэффициент приёма при нормальном слухе, $K_{\text{норм}}$
К3	Нарушенное слуховое восприятие речи	Коэффициент приёма при нарушенном слухе, $K_{\text{нар}}$
К4	Устойчивость к зашумлению	Отношение сигнал/шум, $SNR_{\text{дБ}}$

Все выделенные критерии относятся к типу критериев увеличения, т.е. наилучшим вариантом по каждому из критериев считается тот, у которого значение по данному критерию наибольшее.

На основании всего вышеизложенного, а также с учётом данных табл. 4, можно сформулировать задачу многокритериального выбора в условиях определённости.

Проведение исследования

Для решения поставленной задачи исследования был использован метод ранжирования Борда. Согласно этому методу, выбор наилучшего варианта осуществляется по следующей формуле:

$$R_l = \min_j R_j = \min_j \sum_{i=1}^{n} r_{i,j}\,, j = 1, 2, \dots, m \qquad (5)$$

где $r_{i,j}$ – ранг, который имеет j-й вариант по i-му локальному критерию; R_j – суммарный ранг, который имеет j-й вариант по всем локальным критериям.

С учётом принятых ранее обозначений, формула (5) принимает вид:

$$R_l = \min_j \sum_i r_{i,j}\,, i \in \{K1, K2, K3, K4\}, j = 1, 2, \dots, 6 \qquad (6)$$

Значения локальных критериев сравнения для рассматриваемых вариантов и их ранги представлены в табл. 5. Значения критерия $K1$ получены из табл. 1 и основываются на мнении группы экспертов; значения критериев $K2$ и $K3$ получены из табл. 2 и рассчитываются по формуле (1); значения критерия $K4$ получены из табл. 3 и рассчитываются по формуле (4).

Таблица 5. Значения критериев и ранги сравниваемых вариантов

Критерий / Вариант	К1	К2	К3	К4	Сумма рангов
Бас	0,8 / 4,5	0,8290 / 6	0,8287 / 6	-35,09 / 6	22,5
Баритон	**1,0 / 2**	0,8822 / 5	0,8818 / 5	-34,77 / 5	17
Тенор	**1,0 / 2**	0,9297 / 3	0,9289 / 3	-34,57 / 4	12
Контральто	**1,0 / 2**	0,9032 / 4	0,9030 / 4	-33,14 / 3	13
Меццо-сопрано	0,8 / 4,5	0,9452 / 2	0,9450 / 2	-33,02 / 2	10,5
Сопрано	0,6 / 6	**0,9759 / 1**	**0,9755 / 1**	**-33,00 / 1**	**9**

Наименьшую сумму рангов имеет вариант «сопрано». Следовательно, этот вариант является наилучшим среди рассматриваемых, согласно методу ранжирования Борда.

Заключение

Полученные в ходе исследования результаты свидетельствуют о том, что наилучшим типом голосов для синтезаторов речи является сопрано – женский голос, имеющий самый высокий частотный диапазон. Несмотря на то, что этот вариант занял последнее место по критерию К1 (различимость формант), по всем остальным критериям он занимает первое место. Смещение графика зависимости УЗД от частоты в сторону более высоких частот обеспечило данному варианту, во-первых, высокую степень восприятия как для нормального,

так и для нарушенного слуха (критерии К2 и К3), а во-вторых, высокую устойчивость к зашумлению (критерий К4). Дополнительное изучение значений локальных критериев показывает, что для данного варианта разность значений коэффициентов приёма для нормального и нарушенного слуха является одной из наибольших – это демонстрирует, что вопреки возрастным нарушениям слуха, в большей мере влияющим на данный тип голоса, по сравнению со многими другими типами, его характеристики компенсируют это затруднение.

Хотя выбор конкретного ГБ для программы-синтезатора речи не является целью данного исследования, полученные результаты позволяют в значительной мере уменьшить начальное количество рассматриваемых вариантов, если это потребуется сделать. Дальнейшее сравнение вариантов ГБ внутри класса «Сопрано» можно проводить по различным критериям: техническим, эстетическим и т.д. Кроме того, поскольку данная работа имеет эргономическую направленность (в частности, выразившуюся в выборе локальных критериев сравнения), исследования, имеющие другую направленность, могут продемонстрировать результаты, сильно отличающиеся от представленных.

Библиографический список:

1. Campbell N. Evaluation of Speech Synthesis // Dybkjær L., Hamsen H. and Minker W. (Eds.) Evaluation of text and speech systems. – Springer: The Netherlands, 2007. – P. 29–64.

2. Lemmetty, S. Review of Speech Synthesis Technology. Master’s Thesis, Helsinki University of Technology. – 1999. – 104 p.

3. T. Dutoit. An Introduction to Text-to-Speech Synthesis. – Springer Science & Business Media. – 1997. – 285 с.

4. Бондарко Л. В., Вербицкая Л.А, Гордина М. В. Основы общей фонетики. – 4-е изд., СПб: Академия, 2004, 160с.

5. ГОСТ 12.1.003-2014 Система стандартов безопасности труда (ССБТ). Шум. Общие требования безопасности.

6. ГОСТ Р ИСО 226-2009. Акустика. Стандартные кривые равной громкости.

7. Иванов А. П. Искусство пения. – М.: Голос-пресс, 2006. – С. 37. – 436 с. – ISBN 5-7117-0124-X.

8. СН 2.2.4/2.1.8.562-96. 2.2.4. Физические факторы производственной среды. 2.1.8. Физические факторы окружающей природной среды. Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории жилой застройки. Санитарные нормы (утв. Постановлением Госкомсанэпиднадзора РФ от 31.10.1996 N 36).

9. Соломенник А. И. Структура системы оценки качества синтезированной русской речи // Структурная и прикладная лингвистика. – Вып. 10. – СПб, 2013. – С. 251–266.

10. Соломенник А. И. Технология синтеза речи: история и методология исследований // Вестник Московского университета. Сер. 9. Филология. – 2013. – № 6. – С. 149–162.