ЮГОЗАПАДЕН УНИВЕРСИТЕТ "НЕОФИТ РИЛСКИ"
ПРИРОДО-МАТЕМАТИЧЕСКИ ФАКУЛТЕТ
КАТЕДРА "ИНФОРМАТИКА"

Радослава Станкова Кралева

АКУСТИЧНО-ФОНЕТИЧНО МОДЕЛИРАНЕ ЗА РАЗПОЗНАВАНЕ НА ДЕТСКА РЕЧ НА БЪЛГАРСКИ ЕЗИК

АВТОРЕФЕРАТ

на дисертация за присъждане на
образователна и научна степен "доктор"
в професионално направление
4.6. Информатика и компютърни науки

Научни ръководители:
Доц. д-р Стефан Стефанов
Доц. д-р Борислав Юруков

БЛАГОЕВГРАД
2014

Дисертационният труд съдържа 241 страници, от които 193 страници основен текст и 47 страници приложения. Състои се от въведение, пет глави, заключение, списък на научно-приложните приноси, цитирана литература и седем приложения. Списъкът с използваната литература съдържа 151 източника, от които 17 на кирилица и 134 на латиница.

Дисертационният труд е обсъден и насочен за защита на разширено заседание на катедра „Информатика“ към Природо-математически факултет при Югозападен университет „Неофит Рилски“ – Благоевград, състояло се на 07.10.2014 г.

Дисертантът работи като главен асистент по дисциплините: „Теории, алгоритми и технологии за разпознаване на реч“, „Разработване на приложения за мобилни устройства“, „Мултимедийни бази от данни“, „Практикум по уеб дизайн“, „Програмиране с XML“ и др. в катедра „Информатика“ към Природо-математически факултет при Югозападен университет „Неофит Рилски“ – Благоевград.

Официалната защита на дисертационния труд ще се състои на …….. 2014г. от ……. часа в зала ……..
Защитата ще се проведе пред **научно жури** в състав: 1. Проф. д-р Нина Синягина; 2. Доц. д-р Борислав Юруков; 3. Проф. дтн Райчо Иларионов; 4. Проф. дтн Людмил Даковски; 5. Доц. д-р Йорданка Анастасова.
Резервни членове: 1. Доц. д-р Стефан Стефанов; 2. Доц. д-р Цветозар Георгиев.

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ВЪВЕДЕНИЕ

Възпроизвеждането и разпознаването на реч е естествен процес, който се извършва от всеки човек, още от първия ден на неговото съществуване. Речта е основния начин за общуване, изразяване на чувства и средство за комуникация. Затова е разбираем и интереса на учените към изследването, моделирането и разпознаването на реч. Въпреки напредъка на техниката и технологиите, все още няма толкова усъвършенствани системи, които да са съизмерими с тези човешките способности.

Управлението на работата на говорния тракт и на артикулаторните органи (език, уста и т.н.) е волево, но зависи в голяма степен от допълнителни фактори, като пол, образование, емоционално състояние и др. В резултат на това всеки човек има уникална реч със специфични свойства, като акцент, тон, тембър, бързина. При реалното общуване речта е смесена с фонов шум, ехо, отражение или спонтанни шумове (звън на телефон, шум от телевизор и т.н.). Всички тези фактори превръщат разпознаването на реч в сложна задача.

Разпознаването на реч може да окаже голямо влияние върху онези групи от хора, при които физическите способности са значително намалени. Пример за такива са хората с физически увреждания, малките деца и възрастните. Вграждане на разпознаващи системи в обикновените битови електроуреди, спирки, терминали и т.н. би улеснило много техния живот. Изследванията към настоящата дисертация са насочени към една от тези групи, а именно малките децата на възраст между 4 и 6 години.

Актуалност на проблема

Съвременните технологии за разпознаване на реч са ориентирани основно към речта на възрастни диктори. Детската реч е останала на второ място. Добър пример за приложението на съвременните технологии за разпознаване на детска реч са интерактивните играчки, които могат да разпознаят гласа на детето, изпълнявайки определени команди. Друг пример е гласовата комуникация между компютър и дете. Характерното за всички тях е, че най-често използваният език е английски. Това налага изследване на специфичните акустични и фонетични особености и на други езици, в това число и на българския.

Мотивация

За разработване, обучаване и тестване на системи за разпознаване на детска реч е необходимо наличие на корпус от такава реч на съответния език. Повечето от разработените софтуерни продукти са ориентирани към разпознаване на реч на възрастни потребители. Често използването на тези продукти е неефективно за деца. Това се дължи на редица фактори: Децата могат да заменят или пропускат една или няколко фонеми; Те използват много по-**ограничен речник от възрастните;** Имат богато въображение и асоциативни умения, които използват за създаване на нови собствени думи.

Следователно речникът и говорно-комуникативните способности при децата са твърде различни от тези на възрастните. Затова при създаването на модели за разпознаване на реч трябва да се обърне внимание на техните възможности.

В съвременната литература по разпознаване на реч бяха открити изследвания свързани с английски, италиански, шведски, руски и още няколко други езика, но не бе намерено нищо, свързано с акустичното и фонетично моделиране на детска реч на български език. Ето защо този дисертационен труд е насочен към моделиране на говорима детска реч на български език.

Глава 1 ОСНОВНИ ПОЛОЖЕНИЯ ПРИ РАЗПОЗНАВАНЕТО НА РЕЧ

През последните години в теорията на разпознаване на реч е постигнат напредък по отношение на точността на разпознаване, размера на речника (корпуса), стабилност спрямо околната среда и шумове, възможност за разпознаване на различни стилове на говорене и диалекти. Това е пряко свързано с бързите темпове на развитие на компютърната техника и технологии, възможността за съхраняване и обработка на големи масиви от данни. Довело е до създаването на по-големи корпуси от говорима реч, по-сложни алгоритми и модели за обработка, до постигане на по-голяма ефективност и точност.

1.1 Тенденции в развитие на теорията за разпознаване на реч

Процесът, наречен ***автоматично разпознаване на реч*** (Automatic Speech Recognition - ASR), се състои в извличане на акустичното съдържание (т.е. последователност от думи) от речевата вълна (waveform), записан с помощта на компютърна техника. На фигура 1.1 е представена блок-схема на структурата на система за автоматично разпознаване на реч.

Фигура 1.1: *Основна архитектура на системите за разпознаване на реч*

При продуциране на реч пред микрофон, свързан с компютър, се получава звукова вълна, наречена **речева вълна** (speech waveform), която се преобразува чрез аналогово-цифровия преобразувател от аналогов в цифров сигнал (в модула „**Предварителна обработка на входния сигнал**“). В него се извършва първичното филтриране на сигнала и пречистването му от странични шумове и изкривявания.

При **извличането на характеристични свойства** (feature extraction) се определят **характеристичните вектори** (feature vectors). Те служат за компактно представяне и без загуба на информация на входния сигнал. Има няколко различни набора от функции, които могат да се използват за параметризиране на речевия сигнал. Такива са *Мел-честотен кепстрален коефициент* (Mel-frequency Cepstral Coeficients - MFCCs), *Кодиране с линейно предсказване* (Linear Predictive Coding – LPC) и *Перцептронно линейно предсказване* (Perceptual Linear Prediction - PLP) [3; 57; 80].

Фонетичното моделиране (phonetic modeling), познато още като **моделиране на произношението** (pronunciation modeling) или лексикално моделиране (lexical modeling), е насочено към представяне на базовата характеристична фонетична информация за конкретния език. Това е процес, при който думите се представят чрез графеми на съответните фонеми. С помощта на граф се представя възможната наредба на звуковете за

получаването на смислени думи, налични в речника от произношения. Думите са свързани с вероятностни (възможни) връзки, на които стойността на вероятност е получена при езиковото моделиране. В един език има твърде много думи и би било трудно, и дори немислими да се представят абсолютно всички. Ето защо предизвикателството при това моделиране е да се подберат онези думи, или части от думи, които да удовлетворят представянето на акустичното разнообразие на езика. Също така една и съща дума може да бъде продуцирана от различните диктори по различни начини, което се нарича **коартикулация**, и също трябва да бъде моделирана. Освен това една и съща дума, в частност звук, изказана от един и същи диктор, може да има различно акустично представяне. Съвременни решения на този проблем са предложени в [16; 38; 54].

Акустично моделиране (Acoustic modeling) е статистическо представяне на важни акустични характеристики на основните елементи на речта, обикновено думи или фонеми. Те се основават на използването на общи статистически техники и целят да определят, съвместно с езиковия модел, вероятността за настъпването на дадено лингвистично събитие (например дума, изречение и т.н.), породено от наблюдаваната последователност от характеристични вектори. Два от най-разпространените метода за моделиране на акустичните характеристики на речта са изкуствените невронни мрежи (Artificial Neural Networks) [23; 71] и скритите Марковски модели (Hidden Markov Models - HMMs) [26; 77].

Езиково моделиране (Language modeling) е набор от ограничения, прилагани върху последователността от думи в даден език. Тези ограничения могат да се представят например чрез правилата на пораждаща граматика (generative grammar). За езиковото моделиране най-често се използва така наречения *n-грамен модел* (n-gram model) [68].

При акустично-фонетичното моделиране е необходимо да се направи:

- **Изследване на измененията (variability) в речевия сигнал**: Трябва да се изследва контекстът на говоримата реч, както от фонетична гледна точка, така и от акустична. От фонетична трябва да се обърне внимание на вариациите на един и същи звук (фонема), и неговото различно звучене (фоните). От акустичната гледна точка трябва да се сведе до минимум влиянието на коартикулацията върху процеса на разпознаване. Да се определи вида на разпознаващата система спрямо диктора (зависима, независима или адаптираща се към него) и да се изследва заобикалящата среда (фонов шум, вида на канала за пренос и т.н.);
- **Определят се начините на измерване на грешката при разпознаването**: Прави се сравнение между разпознатия текст и първоначално подадения, като най-общо степента на грешка (Word Error Rate - WER) се измерва, чрез [57]:

$$\text{Степен на сгрешена дума } (WER) = 100\% * \frac{\text{Заместени} + \text{Изтрити} + \text{Вмъкнати думи}}{\text{Броя на думите в правилното изречение}} \qquad (1.1)$$

- **Обработване на аудио сигнала:** Използват се различни методите за извличане на характеристичните вектори, нормализиране на акустичния сигнал и т.н. За установяване на края на речевия сигнал и отстраняване на паузите от тишина се използва подходът end-point detection (установяване на крайната точка), при който се използва спектралния баланс (периода на основния тон) и енергията на сигнала (интензитета) [57]. Акустичните характеристики най-често се извличат с помощта на мел честотните кепстрални коефициенти (MFCCs). За трансформиране на характеристичните вектори може да се използва принципният компонентен анализ (Principal-Component Analysis – PCA), линейният дискриминантен анализ

(Linear Discriminant Analysis) или честотни деформации (frequency warping) за нормализиране на дължината на говорния тракт, които са описани в [30; 92; 111];

- **Фонетично моделиране:** Определят се елементи на речта (фони, фонеми или думи), които са подходящи за моделирани;
- **Акустично моделиране**: Използват се статистически методи за извличане на акустичните характеристики на речта;
- **Използване на адаптиращи алгоритми за намаляване на несъответствията.**

1.2 Речеобразуване и особености на детската реч

Фонаторният апарат е понятие, което се използва за означаване на съвкупността от всички органи на дихателната система. Неговата основна дейност е дишането, но в резултат на еволюцията е придобита вторичната функция **речеобразуване** (гласопроизвеждане). Речта е резултат от изтласкване на въздушен поток под налягане от белите дробове, който преминава през почти затворените гласни струни (гласилките) и тяхното вибриране определя вида на получения звук. Така формираният звук може да рекушира в устната кухина и зъбите, и да се отдели през устата и ноздрите на говорещия.

В резултат от различната позиция на артикулаторните органи се продуцират различни звукове. Резонансите, които са характерни за определена артикулаторна конфигурация, се наричат **форманти**. Формантите се характеризират с главна честота (central frequency), широчина на лентата (bandwidth) и магнитут (magnitude). Формантите са най-важната характеристика, която позволява разпознаването на звуците на речта (voiced sounds). Гласните струни продуцират хармоничната структура, която от своя страна е неподходяща при разграничаването на отделните фонеми [45].

Детската реч е различна от тази на възрастните, по отношение на акустичните и лингвистичните си особености, като периодът на основния тон, интензитета, продължителността на фонемите и честотата на формантите [20; 90]. Тяхната реч се развива с течение на времето, а заедно с това се променят и нейните характеристики. Колкото по-голямо е едно дете, толкова по-устойчиви са измененията, който настъпват във фонаторния апарат. Според [76] експресивната и импресивната реч, периодът на основния тон и честотата на формантите, достигат стойности, характерни за възрастните едва на около 15 години. Това е в контраст с модела на развитие, представен в [49], според който говорният тракт продължава да се променя до навършване на 20 години. От тук се появява несъответствие в различните литературни източници, което все още не е напълно определено.

В ранна детска възраст се наблюдава появата на говорно-комуникативни нарушения, като патологични състояния на гласа (дисфонии), пълната липса на фонация (афония), нарушение на артикулацията (дислалия), нарушения на темпа и ритъма на речта (заекване) и нарушения на писането и четенето (дисграфия, дислекция) [145].

Съвременните изследвания за разпознаване на детска реч са посветени на анализа и извличането на акустичните характеристики на тяхната реч. В [31; 76; 93] са представени различията между акустичните и езиковите характеристики на речта при възрастните и при децата. В [76] е показано, че спектралните и времевите колебания в детската реч са много по-големи спрямо речта на възрастните. Увеличеното разнообразие от получените честоти на форманти с по-голямо застъпване между фонетичните класове при детската реч, спрямо възрастните, прави задачата за класифициране на речта още по-трудно решим проблем. В акустичния анализ на детската реч има значим напредък при определяне на

продължителността на гласните звукове, периода на основния тон и формантите. Все още обаче липсва добре обособен анализ на съгласните звукове [76; 98]. В допълнение трябва да се каже, че почти всички изследвания се отнасят само до американския английски и в малка степен за другите езици като шведски и италиански. Данни за описание на акустичните и фонетични особености на детска реч на български език на този етап липсват.

1.3 Цифрова обработка и анализ на речевите сигнали

Сигналите, които се получават от естествени източници на звук, каквато е човешката реч, се описват чрез непрекъснати функции на времето и затова се наричат **аналогови (непрекъснати) сигнали**. Цифровата обработка на речевия сигнал извършва преобразуването на нивото на електрическо напрежение и дискретизация (sampled) във всеки период от време *T* в цяло число наречено **цифров код** (представено в 16 bits). През определен интервал от време се отчита стойността на получения сигнал, която се изразява с **кодова комбинация**. Това се осъществява от **аналогово-цифровия преобразувател** (analog-to-digital convertor) [149].

1.4 Акустично-фонетично моделиране

В общи линии акустичното моделиране се използва за **моделиране на произношението** (pronunciation modeling), което е свързано главно с определяне на последователността на една или няколко речеви единици (могат да бъдат цели думи, една или няколко фонеми или фони), с цел представяне на големи речникови единици, обект на разпознаването на реч.

В теорията на разпознаване на реч се прави разлика между фонема и фона. Под **фонема** (phoneme) се разбира най-малката звукова единица на речта. Под **звукова единици** или **фона** (phone) се разбира реално продуцираната фонема. За една фонема може да има различни фони, който всъщност се използват за самото разпознаване на съответния фонетичен звук. В повечето случаи фонемите се използва като основни елементи от думата и като набор от модели, представящи разнообразието от различни контексти, в които може да се появи всяка една фонема.

Също така акустичното моделиране включва използването на информация за обратна връзка от разпознаващите алгоритми, за промяна на характеристичните вектори, при наличие на шум или някакви други смущения върху речевия сигнал. Характерно за акустичното моделиране е, че чрез аудиозапис на реч и неговата транскрипция, те трябва да се съпоставят на статистическо представяне на звуците на речта.

1.4.1 Фонетично моделиране

Фонетичното моделиране е насочено към конкретния език и зависи изцяло от него. Ето защо при системи за разпознаване на различни езици е необходимо изграждане на специализирани фонетични модели за всеки отделен език.

За системите с голям речник е трудно изграждането на цял фонетичен модел, тъй като всяка нова задача може да съдържа напълно непознати думи, като например собствени имена или диалектни думи. Друг проблем е, че една и съща дума може да бъде продуцирана по различни начини, както от различни диктори, така и от един и същи диктор.

Основният въпрос, който стои пред фонетичното моделиране е, коя езикова единица да се използва за представяне на фонетичната информация, така че да са удовлетворят следните изисквания [76]:

- **Точност**: Максимално отразяване на езиковото разнообразие и постигане на голяма точност при разпознаването;
- **Обучаемост**: Възможност за използване при обучението на системата за разпознаване и за оценяване на получения резултат.
- **Пораждане**: Избраният елемент трябва да позволява пораждане на нови думи, който не съществуват в модела.

Накратко ще бъдат представени отделните фонетични единици и силните и слабите им страни при използването им в системите за разпознаване на реч.

1.4.2 Акустично моделиране и адаптиране на акустични модели

Добре обучените акустични модели могат да се адаптират към широка гама от променливи. Често обаче се наблюдава известно несъответствие между модела и реалното му изпълнение. Едно възможно решение на този проблем е динамично да се минимизират възможните акустични несъответствия, чрез използване на данни за калибриране (calibration data) и така да се извърши „адаптиране“ на акустичния модел. Техниките за адаптиране могат да бъдат използвани за промяна на системните параметри, за да отговарят по-добре на измененията (вариациите) предизвикани от микрофона, от преносния канал, от шума в околната среда (заобикалящия шум), от диктора и стил (начина) на речеобразуване. Например извършва се адаптиране в зависимост от диктора (speaker-specific adaptation), чрез системи за автоматично разпознаване на реч независима от диктора (speaker-independent (SI) ASR system), с цел създаване на система, зависеща от диктора (speaker-dependent (SD)), но използваща само част от необходимите акустичните данни [41].

Начините за адаптиране мога да бъдат реализирани по няколко различни метода, някои от които са опасани в [32]. Ако транскрипцията на ниво дума (word level transcription) на данните за адаптиране е известна, то адаптирането се нарича **ръководено** (supervised), в противен случай е **неръководено** (unsupervised). При неръководеното адаптиране получения резултат трябва да бъде оценен (estimated). Най-често това се осъществява чрез предварително разпознаване. Адаптирането е **статично** (static), когато всички данни са известни и са зададени предварително. Ако само част от данните за адаптиране са налични и системата продължава да се адаптира с течение на времето, то адаптацията се нарича **динамична** (dynamic).

Три са основни типа адаптиращи алгоритми:

- **Максимално апостериорно адаптиране** (Maximum A-Posteriori - MAP) [43; 58; 84]: Основан е върху теорията на Бейс (Bayes theory) и основната му функция е използването на предварителна информация, получена при процеса на обучение. Така се намалява обема на данни, необходими за получаване на добър акустичен модел, независещ от диктора.
- **Линейна регресия с максимална правдоподобност** (the maximum likelihood linear regression - MLLR) [37; 41]: Най-често се прилага в системи за разпознаване независещи от диктора, които използва скрити Марковски модели с непрекъсната плътност (CDHMMs). Методът MLLR адаптира набор от модели за отделни диктори чрез прилагане на набор от линейни трансформации за Гаусови средни стойности.

Всяка трансформация се използва за няколко Гаусови разпределения. Броят на направените трансформации се определя от размера на данните за адаптиране.

- **Алгоритми за клъстеризация на диктори** (speaker clustering) [135] и eigenvoices (метод на Собствените Гласове - EV) [119]: Методите, базирани на собствените гласове (eigenvoice), са ефективни за бързото адаптиране на диктори, когато е необходимо адаптиране на ограничено количество от данни. В основата на метода е залегнал, принципният компонентен анализ (Principal Component Analysis - PCA), използван основно за намиране на най-важните eigenvoice характеристики.

1.5 Изследвания, свързани с разпознаване на детската реч

За първи път в [132] е направено изследване за взаимовръзката между възрастта на говорещия и точността на разпознаване на реч. В тази разработка говорещите са групирани в пет различни възрастови групи: 8-12 год., 13-18 год., 19-34 год., 35-59 год., и над 60 год. Използвани са езикови модели, които са обучени за всяка отделна езикова група. След това е направен експеримент, като с езиков модел, обучен за определена група се разпознава реч на диктори от друга езикова група. В допълнение са представени постигнати резултати при разпознаване с **възрастово-независим акустичен модел** (age-independent acoustic models), обучен от диктори от различни възрастови групи. По време на експеримента всеки диктор е трябвало да прочете последователност от цифри по телефонна линия. Най-добри резултати са постигнати при наличието на достатъчно количество данни за обучение от съответна възрастова група. Въпреки това резултатите, получени при диктори над 60 год. и на деца от 8 до 12 год., не са удовлетворителни.

Акустичните характеристики на детската реч се променят във възрастта и се очаква, че по-добри резултати ще бъдат постигнати при системи обучени и тествани от деца на съответната възраст. В [51] се използва голям корпус от говорима реч на английски език, на диктори с възрастов диапазон от детската градина до 5 клас, и **групово-определени акустични модели** (group-specific acoustic models). И тук разпознаването е влошено при най-малките деца, но има удовлетворителни резултати при децата от 5 клас нагоре.

Накрая трябва да се отбележи, че през последните години бяха разработени няколко **речево-ориентирани приложения за деца** (speech-oriented applications for children), които използват големи корпуси от детска реч и нововъведенията в технологията за обработка на детска реч. Тези приложения са предимно в областта на четенето с помощта на учители [5; 9; 25; 51; 52; 111] и компютърно подпомогнато обучение по чужди езици [35].

Новата тенденция в технологиите за разпознаване на детска реч е използването и развиването на модели, които са приложими при деца с говорно-комуникативни нарушения. Пример за това е тезис [60], в който се разпознава реч на деца страдащи от дислексия. В [59; 114] са разгледани както проблемите, свързани с увреждания върху говорния тракт, така и с друг важен проблем, който е от съществено значение за степента на разпознаване, а именно наличието на шум в околната среда. В тези източници е постигнато подобряване на получения резултат с 79%.

Методи използвани за адаптиране на детска реч:

- **Нормализиране на дължината на говорния тракт** (vocal tract length normalization - VTLN) [32; 33]: Всеки човек има различна дължина на говорния тракт, а късият

говорен тракт определя по-висок период на основния тон, и обратното. Ето защо при обучение на системите за разпознаване на детска реч от възрастни, в процеса на разпознаване се наблюдава несъответствие с диктора, особено ако това е дете. За компенсиране на тези разлики при извличане на акустични характеристики от речевата вълна може да се използва метода за нормализиране на дължината на говорния тракт (НДГТ). Той се основава на удължаване на тракта с фактор алфа, който мащабира честотния спектър със същия фактор. Например един резонанс от 100 Hz може да се промени на 200 Hz чрез удвояване дължината на тракта.

- **Трансформиране на гласа** (на англ. език voice transformation): В източник [50] е предложено вместо да се извърша НДГТ, да се трансформира речта, веднага след получаването й от микрофона и преди разпознаване. Трансформацията се извършва върху звуковата вълна, при което записаните дискрети на 16 kHz се намаляват до дискрети на 8 kHz. Тъй като детските гласове достигат по-високи стойности на периода на основния тон, то при стандартните записи тези честоти могат да бъдат отрязани. С помощта на този метод се получава пълна информация и запис с високо акустично качество. Въпреки това при използването на трансформиране на гласа се наблюдават и някои странични ефекти, като допълнително ехо.
- **Адаптиране на модел** (на англ. ез. *model adaptation*) [31; 107]: При този метод се извършва адаптиране на детската реч чрез използване на трансформация на параметри, основани на подходящи материали на речта (speech material). Двете ограничения за точността на този метод са размерът на материала за адаптация и начинът на нейното извършване. Използват се помощни материали на речта, които се подбират внимателно, за да не окажат влошаващо въздействие върху разпознаването. Според [31] за адаптирането на реч, записана в голяма зала със значително ехо, е необходимо да се направят много и различни записи на всички диктори. Това позволява използването на един модел за адаптиране на всички диктори и извличане на характеристиките на околната среда. Ако това не бъде изпълнено, ще е необходимо да са разработят отделни акустични модели за приспособяване на речта на всеки диктор и характеристики за всяко помещение, в което е извършен записа. Други техники за адаптиране на модел се основават на разпределението на фонемите по клъстери, в зависимост от техните характеристики, и след това прилагане на една трансформация върху целия клъстер (група). Това е възможно, защото ако две или повече фонеми имат сходни вероятности на плътностите на разпределение, тази зависимост се запазва и след извършване на трансформацията. Формирането на клъстери е полезно от практическа гледна точка, защото данните за адаптиране ще могат да се използват за разпределение на фонемите по клъстери. По този начин може да се осъществи трансформация на фонеми, които не са представени в използвания акустичен модел. Друг критерий за разпределение на фоните по клъстери, който ще осигури трансформирането на непознатите фони в адаптирания материал, е да се групират според сходната им артикулация. Тъй като говорния тракт има еднаква форма за различните възрасти, то може да се твърди, че зависимостта и

характеристиките на тези фони ще се запази и с растежа на децата. Този критерий е използван в дисертацията.

1.6 Цел и задачи на дисертационния труд

От направеното проучване, относно съществуващи модели и алгоритми за акустично-фонетично моделиране на детска реч от една страна, и от друга страна – съществуващите корпуси от говорима детска реч на български език, беше поставена следната цел: **акустично-фонетично моделиране на реч на деца от 4 до 6 годишна възраст и проектиране, създаване и изследване на корпус от говорима детска реч на български език.**

В съответствие с целта се поставят следните **задачи** за изпълнение:

- Да се анализира състоянието на основните положения в съвременната теория за разпознаване на реч;
- Да се проучат особеностите на детската реч;
- Да се анализират фонетичните особености на българския език;
- Да се анализират фонетичните особености на българския език и да се предложи фонетичен модел, отразяващ спецификата му;
- Да се проектира и разработи интерактивна мултимедийна система за записване и анализ на говорима детска реч и обособяването на получените аудио записи в корпус;
- Да се изследват и анализират акустичните характеристики на направените записи на говорима детска реч;
- Да се обобщят получените резултати и да се определи степента на грешка.

1.7 Изводи

Като резултат от изследванията в тази глава могат да се посочат:

- Бяха разгледани по-известните и широко разпространени методи и алгоритми за създаване на разпознаващи системи.
- Бе представен процесът на речеобразуване.
- Бе анализирано развитието на детската реч през отделните възрастови етапи.
- Обърнато бе внимание на необходимите познания и терминологията, използвана в процеса на цифрова обработка на речевия сигнал.
- Проучени бяха добрите практики при фонетичното моделиране и бяха предложени примерни трифонни модели на думи от българския език.
- Бе разгледано значението на околната среда по време на провеждане на интервютата върху резултата от процеса на разпознаване.

Глава 2 ПОСТАНОВКА НА ЗАДАЧАТА ЗА РАЗПОЗНАВАНЕ НА ДЕТСКА РЕЧ

2.1. Етапи при моделиране на детска реч

Етапите, които трябва да се спазват при разработване на системи за автоматично разпознаване на реч, могат да се обобщят по следния начин:

(1) *Събиране на богата база от данни от говорима детска реч, която да отразява богатството на фонетичния състав в българския език.* Отделните звукове и техните съчетания, които трябва да бъдат произнесени, са обособени в езикови единици –

думи или словосъчетания, така че да отговарят на познанията и способностите на малките деца. Записите трябва да бъдат направени в тихо помещение, в домашна обстановка, тъй като от психологическа гледна точка е установено, че детето се чувства най-добре и най-спокойно вкъщи. Така ще се избегне значителното влияние и отклонение, което може да се породи от спонтанността и емоционалното състояние на детето.

(2) *Разработване на контекстно зависимо представяне на фонетичните правила на книжовния български език (фонетичен модел).*

(3) *Извличане на характеристиките на речта.* В потока на свързаната реч е трудно да се установят отделните характеристики на гласните и съгласните звукове и затова се използва най-често Мел-честотните кепстрални коефициенти (Mel-frequency cepstral coefficients).

(4) *Разработване на акустичен модел.* Акустичният модел е фонетичното представяне на говоримата детска реч. Тук се определя съответствието между звук (фона) и буква. Прави се опит да се изведе думата, съответстваща на постъпилата говорима реч.

(5) *Разработването на езиковия модел, който трябва да отразява семантиката и синтаксиса на езика.* Езиков модел на българския език е предложен в [67]. Тук проблемът се състои в това, че децата не винаги се уповават на синтаксиса. Във възрастта между 2-6 години, повечето все още посещават детска градина и не знаят нито да четат, нито да пишат. Изреченията, които използват, могат да бъдат построени по много различни начини.

Това са насоките, които ще се следват при разработването на моделите в настоящата дисертация.

2.2 Диктори

В настоящият тезис под **диктори** ще се има предвид, хората, чиято реч ще бъде записвана, без значение от начина на предизвикване на продуцирането на реч (спонтанна или ръководена). В лингвистичната литература дикторите се наричат още информатори, например в [141; 142].

„Корпусът е един достатъчно обемист лингвистичен архив, в който могат да бъдат открити различни форми на езикова вариативност“ [130] и за постигане на достоверност на събраните данни, е необходимо използване на реч диктори от различни диалектни региони на България [126; 143]. Изследваните диктори са малки деца до предучилищна възраст (до около 6-7 год.).

За да се затвърдят направените наблюдения в чуждата литература, ще бъдат използвани диктори както от мъжки, така и от женски пол, като ще се цели съотношението момче-момиче да е приблизително едно и също. Повече информация за използваните диктори ще бъде представена в Глава 5.

2.3 Корпуси от говорима детска реч

2.3.1 Дефиниция за корпус

Систематично организирана колекция от лингвистични данни, най-често под формата на компютъризирана база от данни, се нарича корпус (corpus в ед.ч. англ. език; corpora в мн.ч. англ. език) [117].

Основните характеристики на съвременните корпуси са:

- Вземане на проби и представяне;
- Краен размер;

- Машинно-четим формат;
- Стандартно представяне.

Корпусите най-често се използват при софтуерната обработка на спонтанна (spontaneous speech) или ръководена (read speech) реч (най-често при четене). Корпусът, свързан с говоримата реч представлява колекция от аудио записи и/или текстове и придружаващата ги транскрипция. В зависимост от начина на организиране на данните в корпусите, те се делят на две категории. Първата категория се нарича корпус от спонтанна реч, в който се съдържа непринудена реч, например породената реч при обикновен разговор. Втората категория се нарича корпус от ръководена реч, и в него се съхраняват данни (реч или текст), които предварително са планирани.

Едно голямо предимство на корпуса от говорима реч (speech corpus) е тяхната практическа полезност при изграждане на софтуерни продукти за обработка на говорима или писмена реч. Те подпомагат технологиите за прилагане на математически базирани набори от статистически данни, наречени акустични модели за всеки отделен звук. В допълнение един такъв корпус би могъл да се използва като база от данни за създаване на интерактивни учители. Също така корпус от говорима реч спомага за изучаването на произношението, словореда и другите езикови модели, характерни за даден език. В [118] е показано, че добре проектирания корпус може да отразява напълно характеристиките на езика, за който е проектиран.

Според [118] **съдържанието на един корпус трябва да бъде избрано не с оглед на езика, а според комуникативната способност на обществото, към което е насочен**. Това правило е използвано при подбора на думите в настоящия корпус, който се основава на първия в България Честотен детски речник. Освен това събраните данни в един корпус разчитат повече на интуицията на разработчика и могат да се приемат за напълно достатъчни за целите на една задача. При разработване на корпус трябва да се анализират събраните данни, да се направят сравнения и изводи.

2.3.2 Съществуващи корпуси от говорима детска реч

След направено проучване на съществуващите корпуси беше установено, че не всички съдържат говорима детска реч. Целта на тази точка е да се представят само онези от тях, които съдържат именно детска реч, и затова останалите корпуси ще бъдат пропуснати:

- Корпусът **CID** е представен подробно в [76]. Материалът за произнасяне (под формата на текстове) е специално подбран, така че да отразява цялото богатство на американски английски;
- Корпусът **CMU Kids** води началото си от 1997 година и негови създатели са Maxine Eskenazi, Jack Mostow и David Graff. Повече информация за този корпус може да се намери на официалния сайт [34; 74];
- Корпусът **CU Children's Audio Speech** [51; 52] е създаден, за да се използва в софтуера Colorado Literacy Tutor, който подпомага гладкото четене при ученици, спомага за придобиване на нови знания, които трудно могат да бъдат усвоени само с четене, и ги насочва към по-лесно писмено изразяване на мисли и идеи;
- Корпусът **PF-Star** [6; 12; 110; 112] е многоезиков корпус от говорима реч на италиански, немски, шведски и британски английски език. Той е създаден по проект, наречен 'Preparing Future Multisensorial Interaction Research' през 2005;
- Корпусът **Tgr-child** [46] съдържа реч на деца от Северна Италия. Записите са били направени в училище. Всяко изречение е показвано на екрана на компютър и след

това детето е трябвало да го прочете. Едно и също изречение, е трябвало да се прочете многократно, до получаване на удовлетворителен резултат;

- Корпусът **ChIMP – Children's Interactive Multimedia Project** [93; 103; 104] съдържа спонтанна реч на 160 момчета и момичета на възраст от 6 до 11 год;
- Корпусът **TIDigits** [78] съдържа говорима реч на английски език, продуцирана от деца и възрастни. Текстът за произнасяне се състои от 22 отделни цифри (по 2 произнасяния на всяка от 11 цифри), 11 двойки от цифри, 11 тройки от цифри, 11 четири последователно наредени цифри и т.н. до 11 на брой седем цифрови последователности;
- Корпусът **Swedish Nice** [9] съдържа реч на деца, записана в училищна и в лабораторна обстановка през 2004-2005 година. Корпусът е разделен на 4 подкорпуса, всеки насочен към определена възрастова група;
- Корпусът **CHILDES (Child Language Data Exchange System)** е многоезикова база от говорима реч и е част от системата TalkBank [123]. Подробно описание за CHILDES може да се намери на официалния сайт [22] и в [86];
- Корпусите **CHILDRU** и **INFANTRU** [85; 146] съдържат говорима детска реч на руски език. на деца. Броят на дикторите в INFRANTRU е 187, а тяхната възраст е от 3 месеца до 3 години;
- Корпусът **ChildIt** [47] съдържа реч при четене на деца от 7 до 13 години.

Като обобщение от настоящото проучване, може да се каже, че **корпус от говорима детска реч на български език не беше открит и е необходимо такъв да бъде създаден**, за провеждане на адекватно изследване и моделиране на фонетичните и акустични особености на детската реч.

На български език съществуват няколко корпуса като BG-SRDat [95], BulTreeBank [18] и т.н, но в тях няма детска говорима реч.

2.4 Фонетично представяне на звуковете в българския език

Фонетиката е науката за звуковото представяне на звуковата система на един език. Основната ѝ задача е описание на фонетичната правила на езика, заедно с начините на артикулация, мястото на учленяването, ударенията на думите, строежа на сричките, интонацията и акустичните свойства на фонемите.

При разпознаване на реч входният аудиосигнал се конвертира в последователност от акустични вектори с фиксирани размери. Нека входния сигнал бъде означен с $Y_{1:M}$, състоящ се от M на брой вектори $y_i, \text{при } i = 1,..M$, и представяне $Y_{1:M} = y_1, y_2, \dots, y_M$. След извличане на характеристичните вектори y_i, се подават към декодиращото устройство (или просто декодер), което се опитва да намери последователността от думи $W_{1:L} = w_1, w_2, \dots, w_L$, съответстваща с най-голяма степен на вероятност на входния сигнал Y. С други думи декодерът се опитва да реши задача от вида:

$$\widehat{W} = arg\max_W\{P(W|Y)\} \quad (2.1)$$

Прякото намиране на вероятността P(W|Y) е трудно за изпълнение. Затова чрез правилото на Бейс (Bayes' Rule), (2.1) се свежда до еквивалентното равенство:

$$\widehat{W} = arg\max_W\{p(Y|W)P(W)\} \quad (2.2)$$

И така най-общо може да се определи, че вероятността $p(Y|W)$ се изчислива чрез **акустичния модел**, а $P(W)$ чрез езиковия (лингвистичен) модел [40].

Речник от произношения (Лексикон): На всяка последователност от думи W акустичният модел съпоставя конкатенация от фонеми, получени в резултат от приложения фонетичен модел (phone models), чрез търсенето на дума, съответстваща на думите в речника от произношения (pronunciation dictionary; pronunciation lexicon).

За да бъде реализирано фонетично моделиране на български език, основано на речник от произношения, е нужно:

- Адекватно фонетично дърво на българския език;
- Вероятностни техники за моделиране за липсващи думи;
- Методи за приближения на думи с променено произношение;
- Да се състави подалгоритъм за транскрибирането на собствени имена, наименования на градове, местности, забележителности и т.н.
- Правилното отразяване на ударението в българския език, фонетичните омофони (думи с еднакво звучене, но с различен буквен състав: шеф-шев; мак-маг; свещ-свеж), графичните омографи (думи с различно звучене, но с еднакъв буквен състав: кàмара-камàра; òси-осѝ).

Според [23] точността на акустичния модел се определя от правилното конструиране на речника на произношенията. Построяването на акустичните правила се извършва ръчно (програмно) и обикновено зависи от спецификата на самия език.

За построяване на речника от произношения и изграждане на фонетичния модел е необходимо да се построи йерархично-фонетично дърво за класификация на фонемите в книжовния български език. **Това дърво трябва да бъде еднозначно определено и ако дадено предположение е правилно за една фонема, то всички останали случаи трябва да са грешни.** За да бъде описаният модел максимално неутрален по отношение на различията в използването на фонетичното представяне, за означаване на звуковете в българския ще се използва нотацията, предложена от International Phonetic Association, а именно Международната фонетична азбука – МФА(International Phonetic Alphabet) [62; 63]. Листата на фонетичното дърво са звуци в българския книжовен език. Въпреки, че в реалното речеобразуване се появяват и други фонеми.

Фонетичната транскрипция се изгражда основавайки се на точното съответствие между буква и звук. Най-малката различима единица се нарича **фонема**. Всяка фонема може да има различни нюанси на произношение, зависещи от диктора, които се наричат **фони** (phone) или още **говорни звукове**. Говорните звукове, които са фонетично подобни, но се определят от позицията си в думата, се наричат **алфони**. За разлика от говорните звукове, алофоните представляват една и съща фонема, като могат да се заместват една друга, в зависимост от фонологичните условия [139].

При формалното представяне на естествения език и неговата граматика, обикновено алгоритмите са независими от конкретиката на езика. По друг начин нещата стоят за изграждането на фонетичното дърво. Тук трябва да се вземат предвид редица допълнителни фактори:

- Вида на гласните, начина на артикулирането и позициите, които могат да заемат;
- Вида на съгласните, начина на артикулиране и позициите, които могат да заемат;
- Потъмняване на гласни;
- Изпадане на съгласни;
- Редуциране на звукове;
- Глайдове;

- Ударения.

Като обобщение може да се каже, че при изследване на фонетичното богатство на даден език е необходимо да се разгледат всички алофони, да се отразят зависимостите между тях и да се установят условията на тяхното проявление. Това ще бъде направено в Глава 4, където ще се изградят формални правила и дървета за решения за всяка фонема, отразяващи влиянието на съседите й. Предложеният фонетичен модел ще се базира на източниците [137; 139; 144].

Във фонетичния строеж на българския език гласните (означени с V) се съчетават предимно със съгласни (означени със C), т.е. [VC] или [CV]. Срещат се съчетания от вида гласна-гласна, където $[V_1V_2]$ при $V_1 \neq V_2$ или $[V_1V_1], \text{при } V_1 = V_1$ (табл. 2.3). Вариантът за съчетаване на две еднакви гласни се използва само за чуждиците (думи с небългарски произход). Оттук следва, че всяка гласна може да се съчетава с всяка гласна.

Когато настъпва комбинаторна промяна между два звука от един и същи вид, т.е. между две гласни $(\mathrm{V_iV_j}, \text{при } \mathrm{V_i} \neq \mathrm{V_j})$ или две съгласни $(\mathrm{C_iC_j}, \text{при } \mathrm{C_i} \neq \mathrm{C_j})$, тогава тя се нарича асимилация (уподобяване).

2.5 Обосновка на метода за акустично моделиране

В резултат от направените изследвания в първа глава на методите за разпознаване на детска реч бе установено, че най-широко е използван похватът свързан с адаптирането на речта към модел разработен за възрастни диктори. Това има редица ограничения, свързани с допълнителни изчисления и забавяне на процеса по разпознаване. Беше показано, че получените резултати при тези подходи често не са удовлетворителни при малки деца.

Ето защо в настоящия тезис се използва подход основан изцяло на спецификата на детската реч. За целта ще бъдат изследвани акустичните характеристики на отделните диктори от различни възрастови групи. Получените данни ще бъдат използвани като входни параметри на модифицирания алгоритъм за клъстеризация ИСОМАД. Резултатите ще бъдат оценени чрез използване на метода за определяне на степента на грешка и са представени подробно в четвърта и пета глава.

2.6 Архитектура на системите за автоматично разпознаване на реч

Представена е архитектура на диалогова система за разпознаване на реч, при която потребителят комуникира с разпознаващото устройство чрез говорима реч. Този тип системи обикновено се състои от пет основни модула, които са проектирани да работят заедно. Модулите *„Езиково моделиране“* и *„Диалогов мениджър“*, веднъж разработени, могат да се използват и за други системи, или да бъдат подменени с напълно нови такива. Модулът *„Предварителна обработка на входния сигнал“* включва филтри за премахване на шум от входния сигнал. След това аналоговият сигнал се преобразува в цифров с помощта на аналогово/цифровия преобразувател. В блока *„Разпознаване на реч“* се извършва извличането на всички характеристични вектори от цифровия поток от данни, които впоследствие се декодират, т.е. се разпределят (клъстеризират) векторите по класове (клъстери). Този блок е свързан с базата от данни от говорима реч (наречен още *корпус от говорима реч*), която е използвана както за обучение, така и е необходима за самоорганизиране и разпознаване на новопостъпилата реч, било то отделни думи или цели изречения. След което речта се преобразува в текст, т.е. се извършва разпознаване.

2.7 Спецификация на софтуерната разработка на мултимедийна система за създаване, управление и анализиране на корпус от говорима детска реч

Цели и задачи: Като очаквани резултати могат да се посочат удобен графичен потребителски интерфейс, с лесна достъпност до всички функции и данни, с обработени всички възможни изключения, които могат да възникнат при работа със софтуера, така че потребителят да е наясно във всеки един етап какво се случва и как да реши възникнал проблем.

Технически параметри при разработка: Разработване на ориентирани към потребителя външен вид и система за управление и навигация, организирана под формата на менюта и бутони с подсказващи съобщения; Базата от данни трябва да съдържа всички необходими данни, които са пълна информация за всеки един диктор (име, адрес, снимка, учебно заведение, семейство, увреждания на говорно-комуникативния апарат и т.н), данни за населени места, данни за учебни заведения, данни за възможни болести и т.н.

2.8 Изводи

В тази бе направено:

- Определяне на етапите за моделиране на детска реч;
- Дефинирани бяха понятия като корпус и диктор;
- Представени са някои от съществуващите съвременни корпуси от говорима детска реч и корпуси от говорима реч на български език;
- Обобщени са техните положителни и отрицателни черти, езиковото им богатство и големината им;
- Бяха разгледани фонетичните особености на българския език;
- Направена бе схематично представяне на архитектура на цялостна система за разпознаване на реч;
- Изготвена бе спецификация за софтуерна разработка на мултимедийна информационна система за управлението на корпус от говорима детска реч;
- От направеното проучване е установено, че към момента не съществува база от данни със записан детска реч на български език и това представлява основна пречка за разработване на разпознаваща система, или прилагането на вече съществуваща;
- Поради тази причина преди извършване на акустично-фонечитното моделиране на говорима детска реч, е необходимо създаването на такъв корпус и разработване на софтуерен продукт, с дружелюбен потребителски интерфейс, предназначен за неговото управление.

Глава 3 ПРОЕКТИРАНЕ НА КОРПУС ОТ ГОВОРИМА ДЕТСКА РЕЧ НА БЪЛГАРСКИ ЕЗИК И РАЗРАБОТВАНЕ НА МУЛТИМЕДИЙНА СИСТЕМА ЗА РАБОТА С НЕГО

В практиката събирането на говорима реч от деца се е оказала трудоемка задача и се е наложило използването на допълнителни средства. Например чете се думата, която трябва да бъде произнесена, и след това детето я повтаря [14]. Често се случва подбраните думи да бъдат твърде сложни и детето да срещне трудност при тяхното изговаряне. Освен това малките деца лесно губят концентрация и се разсейват, което още повече затруднява

събирането на данни. Във връзка с това е необходимо да бъде създадена база от данни, която да позволява записване на допълнителна информация, свързана с подпомагане процеса на запълване на речева база от данни с говорима реч.

Необходимо е обединяването на цялата мултимедийна информация, съпровождаща записването и обработването на говорима реч, и улесняване работата на изследователите в областта на обработването на реч. Тази база от данни е основополагаща за разработването на системи за разпознаване на детска реч и в бъдеще на системи за синтезиране на детска реч. Освен това корпусът може да бъде използван от логопеди за създаване на пълен профил на изследваните от тях деца, както и извеждане на статистическа информация.

Проектирането и разработването на корпус от говорима реч е неделима част от системите за разпознаване на реч. Тяхното качество и обхватът до голяма степен афектират върху дейността на разпознаващите устройства. Ето защо всяка база от данни трябва да отразява фонетично богатство на изследвания език.

3.1 Анализ на проблема

Според [1] корпусите от говорима детска реч биват разработвани поради две основни причини – първата е за провеждане на фундаментални научни изследвания на акустичните, фонетичните, лексикалните, семантичните и синтактичните прояви на даден език; и втората е, за установяване на различията между отделните диктори, като пол, възраст, заобикаляща среда, канали за пренос на данни и т.н.

Освен това, както бе показано в Глава 1, речта на децата съдържа набор от конкретни параметри [76], което ги превръща в група от потребители със специфични изисквания към системите за автоматично разпознаване на реч. Често срещаните говорни нарушения са друг специфичен проблем за разрешаване [27]. В момента основна тенденция при колекционирането на данни от говорима реч (speech data) на деца от 4 до 6 години в момента е анализирането на акустичните и езиковите характеристики.

При провеждане на проучването, към момента на писане на тази дисертация, не бе открита корпус от говорима детска реч на български език.

3.2 Подготовка за разработване на интерактивна мултимедийна система за управление и анализ на корпус от говорима детска реч

От техническа гледна точка корпусът е своеобразна база от данни, за съхраняване на информация под формата на текстове и/или аудиофайлове. Ето защо в настоящия раздел ще бъдат представени трите основни аспекти, които трябва да се спазват при проектиране на база от данни от говорима реч: видът на използвания речник, броят на проведените сесии с един диктор и техническите аспекти на записа [72].

3.2.1 Речник

Според [50] за постигане на удовлетворителни резултати при разпознаването, обучението (тренирането) на разпознаваща система, трябва да бъде извършено от същата целева група, чиято реч ще се разпознава. Поради тази причина записите трябва да са ориентирани към малки деца (от 4 до 6 год.) и да се предоставя възможност за бърз и удобен запис, като се спазва баланса между пола и възрастта им. Обикновено речниците обхващат пълния набор от фонеми на съответния език. Тук се появява първата трудност свързана с малките деца, която е основана на различния от възрастните речник. Поради тази причина при настоящото изследване, за основа се използва детският честотен речник

предложен в [138]. Той съдържа най-често използваните думи от децата на възраст между 3 и 7 години, както в тяхното ежедневие, така и в учебните пособия и помагала, които се използват в детските учебни заведения. Посочени са думите и честотата на тяхната употреба. Чрез признака **ранг на думата** се посочва степента на значение на дадена дума в детските текстове.

3.2.2 Брой сесии

Използваните диктори трябва да обхващат деца на възраст от 4 год. до 6 год. Под сесия ще се разбира един запис (файл), направен от един диктор в определен период от време. Тук се определя продължителността на направените записи и броя на сесиите с всяко дете, което участва в изследването. Записите трябва да са максимално кратки, защото децата на тази възраст бързо се уморяват. Всеки запис ще съдържа произнесена реч само от един диктор. Поради тези причини думите, които трябва да каже едно дете, за една сесия трябва да са ограничени.

3.2.3 Технически аспекти

Към техническите аспекти спадат представяне на околната среда, техническите средства, с които ще се извършва записа (звукова карта, вид на микрофона и т.н.) и методи за изчистване на сигнала от шум. Речта трябва да бъде събрана по реалистичен начин. Това означава, че записите не трябва да бъдат направени в звукозаписно студио, което ще отличава получения корпус от повечето съществуващи колекции от записи на говорима детска реч.

От психологическа гледна точка е установено, че децата в целевата възраст се чувстват най-спокойно и могат по-лесно да комуникират в позната обстановка, например като техният дом, учебното заведение (детска градина или училище), в което се обучават и др. Това е и средата, която е избрана за направата на аудиозаписите. Всеки записи ще съдържат последователност от думи, които предварително ще бъдат обособени като колекции от думи. Предвидени са и записи от свободен разговор, в които детето може да беседва с провеждащия експеримента или да говори само.

3.3 Концептуален модел на интерактивна мултимедийна система за управление и анализ на корпус от говорима детска реч

Изграждането на концептуален моден на софтуерен продукт, който да отговаря на изискванията: да съдържа освен думите и записите от говорима реч, и мултимедийни обекти, да подпомага назоваването на съответната дума от дикторите, да позволява обработването и анализирането на получените записи, е трудоемка задача.

В настоящия раздел ще се представят характеристиките, отличаващи предложения модел от съществуващите корпуси от говорима реч. Както бе показано корпусът е база от данни, в която се съхраняват аудиофайлове, текстът, който е произнесен, и неговата фонетична транскрипция. Разликата между тривиалните корпуси е, че в ChildBG се събират и допълнителни данни, които могат да допринесат за изграждането на цялостен профил на диктора, неговите специфични особености, влиянието на околната среда, върху качеството и начина на продуциране на реч, както и психологичното състояние на диктора по време на самия запис.

Всяка дума в ChildBG се представя чрез буквена и фонетична нотация. За фонетичното представяне на думите се използва машинночетимата фонетична азбука X-SAMPA (Extended Speech Assessment Methods Phonetic Alphabet) [130]. Всяка дума е онагледена с

изображение (едно или няколко) и/или звуков файл (един или няколко). Думите са представени и като части на речта. Фонетичното представяне е резултат от предложения модел в Глава 4.

Думите са организирани в колекции от думи, които съдържат от 2 до 32 думи. Предвидени са колекции за спонтанна реч, в които няма нито една дума. Съвкупността от думи в една колекция се определя от изследователя, а дикторът по време на запис ще може да вижда изображенията или да чува асоциирания с тях звук.

Една сесия ще се провежда върху една колекция от думи. Изследователят поставя началото на сесията и определя, кога да започне и да спре записването на продуцираната реч от диктора. Така ще се получи един запис. Процесът на извършване на записването на реч се нарича интервю. Под сесия ще се разбира един запис на реч на един диктор върху една колекция от думи.

Впоследствие записите трябва да бъдат обработени и представени като отделни думи. Тези подзаписи ще се наричат обработени записи.

В корпуса ChildBG ще може да се съхраняват допълнителни данни за диктора, като: възраст; пол; данни за грижещия се за него (когато диктора е дете); данни за посещаваното учебно заведение (ако има такова); данни за посещаване на допълнителни курсове; и данни за отклонения (ако има такива) в говорно-комуникативните способности на детето. Тези разширени възможности отличават предложения модел от стандартните корпуси, разгледани в Глава 2. Това е схематично представено на фигура 3.1.

Фигура 3.1: *Представяне на разликите между стандартните корпуси от говорима реч и корпус ChildBG.*

От горната фигура се вижда, че всички характеристики, които оказват въздействие върху продуцирането на реч, като емоционално състояние, околна среда и т.н., могат да се запише и използва в последствие от ChildBG.

По време на записа детето трябва да назовава дума по представени изображения и по чутите звукови файлове. Възможно е детето да каже съвсем различна дума от очакваната, тогава думата се маркира като спонтанна реч. В краен случай, ако изпита затруднение и не може да разпознае обекта или действието, то ръководителят може да се намеси, подпомагайки допълнително детето, като назове думата, или му зададе насочващ въпрос. Повечето системи от този тип не позволяват използването на интерактивни елементи по

време на запис на говорима реч, като колекция от различни изображения или аудио звуци. В настоящата система е отстранен този недостатък.

За да се постигне всичко това, е необходимо данните да бъдат събрани в релационна база от данни. Тъй като идеологията на релационните бази от данни не е цел на настоящия дисертационен труд, теорията за това няма да бъде подробно разгледана.

Предложената схема на релационна база от данни е нормализирана. Всички същности имат първичен ключ, нямат многостойностни атрибути, всички данни са атомарни (неделими), нямат частични и транзитивни функционални зависимости (*Приложение 3)*. Трите основни асоциативни същностни, върху които се акцентира, са: речникът (таблицата Words), дикторите (таблицата Speakers) и направените записи (таблицата Records).

Информацията, съхранена за всеки един от дикторите, се състои от: трите имена, пол, възраст (години и месеци), дата на раждане, текущ адрес, брой на деца в семейството, номер на детето по реда на раждане, посещава ли детска градина, и ако да – коя, посещава ли допълнителни курсове и какви (логопед, уроци по пеене, уроци по музика), отклонения от нормалното развитие и болести.

Информацията, която ще се съхранява за направените записи, е: място на записа, използвана апаратура, емоционално състояние на диктора (детето) по време на записа и характеристиките на заобикалящата среда. Получените файлове ще са с ниска компресия и във формата WAV. Те могат да съдържат външни шумове, като гласа на майката или ръководителя на експеримента (наречен накратко експерт), неречеви звуци, издадени от диктора и т.н. Аудиофайловете с много смущения ще бъдат отстранени от корпуса.

3.4 Софтуерна реализация на интерактивна мултимедийна система за управление и анализ на корпус от говорима детска реч ChildBG

В този раздел ще бъде представена програмната реализация на интерфейса и работата с интерактивната мултимедийна система за работа с корпуса ChildBG. Софтуерът притежава графичен потребителски интерфейс и е разработен с безплатната среда Turbo C++ Explorer. Базата от данни е създадена чрез системата за управление на релационни бази от данни (СУБД) Microsoft Access, но е експортирана и във формат за Microsoft SQL Server.

3.4.1 Архитектура на софтуера за работа с корпуса ChildBG

Интерактивната мултимедийна система за работа с корпус ChildBG е софтуер, който позволява записването и използването на говорима детска реч, поддържане на електронни профили на дикторите, информация свързана с условията, при които са проведени записите, както и емоционалното и психическо състояние на дикторите. Първите резултати за разпознаване на емоции при децата в предучилищна възраст са представени в [67].

Корпусът ChildBG е основополагащ за изграждане на модели за разпознаване, обработване и използване на детска реч.

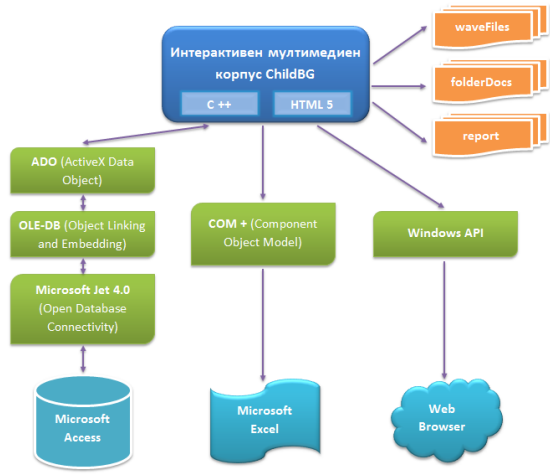
При разработването на такъв софтуер трябва да бъдат решени следните задачи [118]:

- Да се определи операционната система, на която ще бъде инсталиран;
- Да се определи развойната среда за разработване, която ще се използва за неговото създаване;
- Да се определят всички необходими данни и да се реализира релационната схема на базата от данни;

- Да се определи стилът на интерфейса, като цветова палитра, бутони, изображения и т.н. За целта най-често се използват ескизи.
- Да се определи видът на отчетите, които ще се генерират;
- Да се определи физическата организация на приложението след инсталация върху твърдия диск;
- Да се определи начинът на разпространение и цена.

Първите три от посочените по-горе изисквания бяха описани в Глава 2. Останалите четири задачи, ще бъдат решени в настоящия раздел.

Под физическа организация ще се разбира реалната архитектура на приложението, заедно с всичките му необходими подпапки (waveFiles, forlderDocs, report), доставчиците на услуги (ADO, OLE-DB, Microsoft Jet 4.0, COM +) и хранилищата за данни (Microsoft Access или Microsoft SQL Server), както и на други външни приложения (Microsoft Excel и Web Browser), с които системата може да взаимодейства (фиг. 3.4).



Фигура 3.4: *Схематично представяне на физическата организация на софтуерния продукт за работа с ChildBG*

При определяне на интерфейса на софтуерния продукт от съществено значение е удобството и тематичната организация на отделните прозорци. Използва се диалогово ориентирана организация, при която от главния прозорец на приложението посредством менюта се достъпват всички останали прозорци. Менютата са организирани по теми: „Основно“, „Речник“, „Диктор“, „Корпус от говорима реч“ и „Помощ“ (фиг. 3.5).

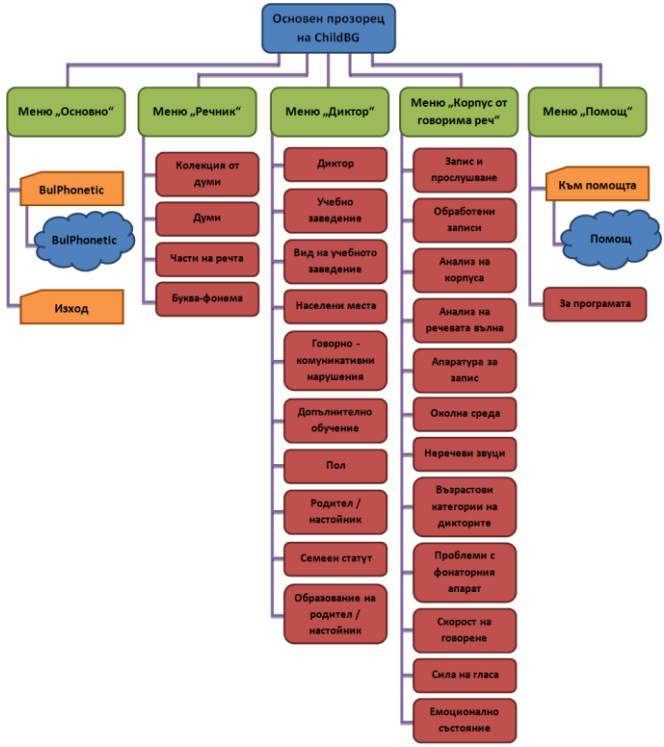
От **меню „Основно“** може да се стартира уеб приложението BulPhonetic [16] и да се затвори цялото приложение чрез командата „Изход“. Уеб приложението “BulPhonetic” предоставя възможност за автоматичното транскрибиране на текст на български език на няколко азбуки за фонетично представяне.

В **менюто „Речник“** тематично са обособени всички функционалности, които са необходими за изграждане на речника от думи в един корпус. Оттук могат да се стартират прозорците: „Колекции от думи“, „Думи“, „Части на речта“, „Буква-фонема“, като се използват едноименните команди.

От **меню „Диктор“**, отново чрез едноименни команди, могат да се стартират прозорците: „Диктори“, „Учебни заведения“, „Вид на учебното заведение“, „Населени места“, „Говорно-комуникативни нарушения“, „Допълнително обучение“, „Пол“, „Родител/настойник“, „Семеен статут“ и „Образование на родителя/настойника“.

В **меню „Корпус от говорима реч“** са обособени всички необходими команди, от които могат да се стартират прозорци, позволяващи записване на говорима реч и осигурява достъп до останалите прозорци, разширяващи основната функционалност. Оттук чрез едноименни команди могат да се стартират прозорците „Запис и прослушване“, „Обработени записи“, „Анализ на корпуса“, „Анализ на речевата вълна“, „Апаратура за запис“, „Околна среда“, „Неречеви звуци“, „Възрастова категория на дикторите“, „Проблеми в елементи на фонаторния апарат“, „Скорост на говорене“, „Сила на гласа“ и „Емоционално състояние“.

От **меню „Помощ“** чрез командата „Към помощта“ може да се стартира помощната система на приложението, за която е необходимо наличие на уеб браузър, тъй като използваните файлове са в html формат. Освен това оттук чрез командата „За програмата“ се стартира едноименният прозорец, съдържащ информация за неговия разработчик.



***Фигура 3.5**: Схема на архитектурата на интерфейса на софтуерния продукт за работа с ChildBG*

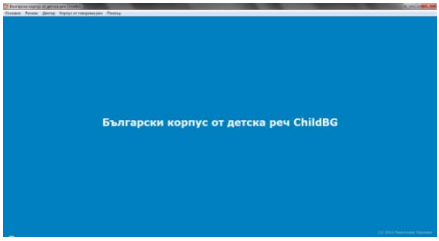
За реализирането на всички отчети в ChildBG се използва технологията COM+ (Component Object Model), с помощта на която се извършва експортиране на данни към

Microsoft Excel. Изключение прави само отчетът, съдържащ данни за думите и техните изображения, който е разработен с езиците HTML и CSS, и визуализиран в уеб браузър

3.4.2 Функционални възможности на интерактивната мултимедийна система за работа с корпуса ChildBG

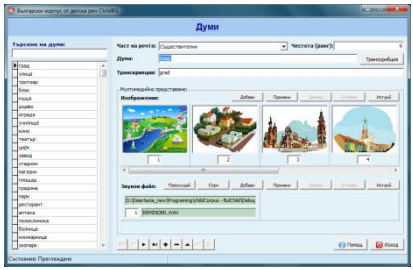
Софтуерния продукт за работа с корпуса ChildBG предоставя много повече функционални възможности от стандартните корпуси за говорима реч, пример за такива са CHILDRU [79], Emu Speech Database System [15] и др. Голяма част от корпусите от говорима детска реч са разгледани в раздел 2.3 на втора глава. Характерното за повечето такива корпуси е недружелюбният интерфейс и несъобразността с възможностите на децата между 4-6 години за продуциране на реч. В настоящия раздел ще бъдат представени функционалните възможности на софтуерния продукт, който носи същото наименование като корпуса, ChildBG.

След стартиране на интерактивната мултимедийна система за работа с корпус от говорима детска реч ChildBG се зарежда основният прозорец (фиг. 3.6). Той е стандартен прозорец и притежава всички системни бутони за максимизиране, минимизиране и затваряне на приложението. Тук се намира главното меню, от което са достъпни всички останали прозорци.



Фигура 3.6: *Основен прозорец от интерактивната мултимедийна система за работа с корпуса ChildBG.*

Друг важен прозорец от софтуерния продукт, който ще бъде разгледан, е прозорецът „Думи“ (фиг. 3.9). Той съдържа целия речник, който може да се използва при записване на реч. За всяка дума може да се определят следните данни: Каква е като част на речта; Честота й на срещане; Нейната транскрипция съгласно фонетичната азбука XSAMPA, се извежда автоматично при натискане на бутона „Транскрипция“.



Фигура 3.9*: Прозорец “Думи”*

Към всяка дума потребителят може да добави изображения (в BMP формат) и аудио файлове (в WAV формат). Изображенията се записват в базата, а аудиофайловете се прехвърлят в папката folderDocs, която се намира в същата директория, в която е изпълнимия файл на приложението. За разлика от повечето корпуси, речника може да се разширява и допълва с думи, което също е предимство за предложения модел на ChildBG.

За да се улесни процесът на записване, всички думи са организирани в колекции от думи. Има възможност една и съща дума да се среща в повече от една колекции. По този начин може да създадат различни видове колекция, с почти едно и също съдържание на думи, за различните възрастови групи.

***Фигура 3.10**: Прозорец "Колекции от думи"*

Тази функционалност е достъпна от прозорец „Колекции от думи" (фиг. 3.10). Цялата област на диалоговия прозорец е разделена на три части. В левия панел може да се търсят и избират думи, които да бъдат добавени посредством бутона „стрелка" към текущата колекция. Всички избрани думи са видими в средния панел. Тук се определя името на колекцията. Необходимо е да се даде уникална абревиатура на колекцията, която по-късно ще бъде използвана при формиране на името на записвания аудио файл (табл. 3.1).

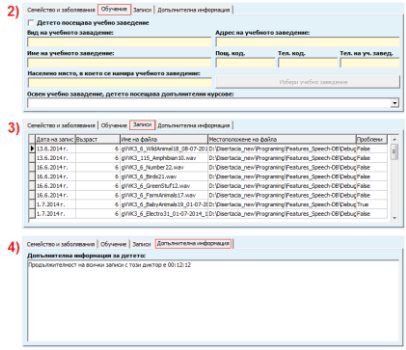
В десния панел има възможност да се видят всички имена на наличните колекции от думи, като при избор на някоя от тях, тя се активира и данните й се визуализират в средния панел.

От този прозорец чрез бутони намиращи се в десния панел, могат да се генерират три справки към Microsoft Excel. Те са свързани с анализ на буквите в текущата колекция, анализ на буквите във всички колекции и анализ на използваните думи колекциите. От бутона „Отпечатай" се генерира справка, под формата на уеб страница, използвана за картотекиране на колекцията.

Една от най-важните функционалности на настоящия софтуер е възможността за поддържане на електронен профил на всеки един от дикторите. Тази функция е достъпна от прозореца „Диктори" (фиг. 3.21 и фиг. 3.22).

Този прозорец позволява търсене на диктор по име. Освен това тук се поддържат всички негови данни, така че да може да се изгради правилен психологичен и лингвистичен профил. Всеки въведен диктор има уникално системно име, определено полуавтоматично, тъй като потребителят сам избира каква ще е абревиатурата, която ще обозначава диктора, а системата с помощта на бутон дописва уникален код, гарантирайки, че няма да има двама диктори с едно и също системно име.На табл. 3.2 е представен пример за създаване на уникален идентификатор на един диктор. Името се формира като поредица от буквата b (boy) за момче или g (girl) за момиче, буквена последователност от името на детето (Виктория Велинова Кралева -> VVK) и уникален номер (първичния ключ на записа на диктора, взет от базата от данни. Така се получава означението gVVK3, което се използва за формиране на имената на аудио файловете, създавани при провеждане на интервютата с диктора.

***Фигура 3.21**: Прозорец "Диктори", с видима страница „Семейство и заболявания".*



***Фигура 3.22**: Части от прозореца "Диктори", с представяне на отделните страници: 2) „Обучение", 3) „Записи" и 4) „Допълнителна информация".*

Информацията за всички диктори може да се експортира под формата на отчет (справка) към Microsoft Excel, с помощта на бутона "Експорт към Excel".

Тъй като софтуерът е насочен към събирането на реч на малки деца, на страницата „Семейство и заболявания" са отразени важни фактори, оказващи въздействие върху продуцирането на реч, като семеен статус на родителите, броя на децата в семейството, кое по ред дете е дикторът, има ли говорно-комуникативни отклонения, като например дислексия, заекване и т.н. Заболяванията и отклоненията на дикторите могат да се изберат от прозореца „Говорно-комуникативни нарушения". Всички допълнителни данни, необходими за изграждане на профила, се управляват от потребителя с помощта на допълнителни прозорци, които ще бъдат описани по-нататък в този раздел. За попълване на информацията на тази страница се използват данните, въведени в прозорците „Населени места", „Говорно-комуникативни нарушения", „Пол", „Родител/настойник", „Семеен статут" и „Образование на родителя/настойника".

От страница „Обучение" на прозореца „Диктори" (фиг. 3.22: 2) може да се въведат данни за посещаваното от детето учебно заведени (ако има такова), както и данни за допълнителни курсове (ако детето посещава такива). За попълване на информацията се използват данни, въведени в прозорците „Учебни заведения", „Видове учебни заведения" и „Допълнително обучение".

От страница „Записи" на прозореца „Диктори" потребителят на системата може да види всички налични записи свързани с настоящия диктор, възрастта му по време на тези записи, колекциите от думи, върху които са направени записите, тяхната продължителност и т.н. (фиг. 3.22: 3). Тук се използва информация от прозореца „Записи и прослушване на говорима реч", която е достъпна от менюто „Корпус от говорима реч".

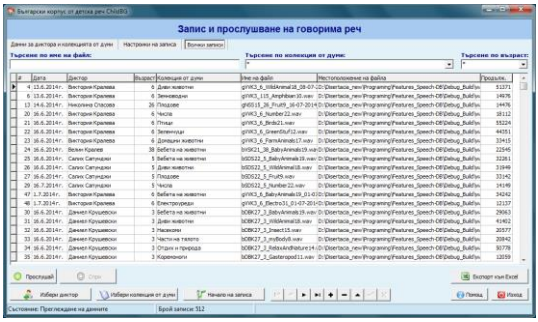
На страница „Допълнителна информация" на прозореца „Диктори" е налично многоредово текстово поле, в което могат да бъдат въведени данни по избор на потребителя. На фиг. 3.22: 4) е показан пример за въведена обща продължителност на всички записи, направени от избрания диктор.

Прозорецът „Анализ на корпуса" (фиг. 3.32) позволява да се анализират всички налични данни в корпуса. След определяне на съответните критерии, по които да се

филтрират данните, те могат да бъдат експортирани към отчет (справка) на Microsoft Excel. Възможностите за анализ са обособени в три отделни страници: „Диктори“, „Записи“ и „Обработени записи“. На всяка една от страниците филтрирането се извършва посредством падащи списъчни полета. Бутонът „Без филтър“, който е наличен във всички страници, изчиства избраните стойности в полетата, при което се визуализират всички налични данни.

Най-важният прозорец от разглеждания софтуерен продукт ChildBG е **прозорецът „Запис и прослушване на говорима реч“**. Той служи за създаване, редактиране и изтриване на проведените сесии с дикторите. В него се описва цялата необходима информация за направения запис, обхващаща емоционалното състояние на диктора, заобикалящата го среда, проблеми с фонаторния апарат и т.н. Интерфейсът на прозореца е организиран под формата на три взаимно свързани страници.

Първа страница се нарича **„Всички записи“** (фиг. 3.36) и предоставя възможност за преглед на всички направени записи на говорима реч. Наличните данни могат да се филтрират по име на файл, колекция от думи и възраст на дикторите. Тук се намират бутоните за прослушване / спиране на записаната говорима реч. Освен това от тази страница, чрез бутона „Експорт към Excel“, може да се генерира справка, която да се експортира към Microsoft Excel, съдържаща обобщение на всички данни за записите. Тук се намират и бутоните „Прослушай“ за прослушване и „Спри“ за спиране на избрания запис



Фигура 3.36: *Страница „Всички записи“ от прозорец “Запис на говорима реч”.*

От **страницата „Данни за диктора и колекцията от думи“** на прозореца „Запис и прослушване на говорима реч“ (фиг. 3.37) се преглеждат данните за диктора, чиято реч е записана (или ще се записва), и се преглежда (или избира) колекцията от думи, която е използвана (или ще се използва) в записа.

От **страницата „Настройки на записа“** (фиг. 3.38) се предоставя възможност на потребителя да определи данните за условията на околната среда, възрастовата категория на диктора, емоционалното състояние на диктора, скорост на говорене и силата на гласа на диктора, наличието на неречеви звуци [Смях; Шум от вентилатор; Падане на предмет; и т.н], проблеми с елементи на фонаторния апарат [Липса на зъб/и; Запушен нос; Ларингит; и т.н.], участвал ли е експертът по време на записа, диктора срещнал ли е трудности при речеобразуването, записът съдържа ли спонтанна реч и др. Също така може да се определи хардуерът и софтуерът използвани за провеждане на записите. Всички тези данни трябва предварително да бъдат въведени чрез прозорци „Апаратура за запис“,

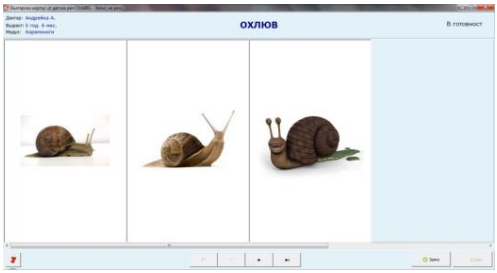
„Околна среда“, „Неречеви звуци“, „Възрастова категория на дикторите“, „Проблеми в елементи на фонаторния апарат“, „Скорост на говорене“, „Сила на гласа“ и „Емоционално състояние“.

Възможно е да се определят характеристиките на получения аудиофайл: честотата на дискретизация (по подразбиране е избрана 44100 Hz), битовете за дискрет (по подразбиране е избрано 16 bit) и броя на каналите (по подразбиране е избран „Моно“, тъй като „Стерео“ аудиофайловете по-трудно се обработват при разпознаването на реч). Потребителят има възможност при необходимост да въведат и допълнителна информация в многоредово текстово поле. Честотата на дискретизация (sample rate) показва броя на дискретите за една минута. Всички получени файлове са в некомпресиран вид и са във файлов формат WAV. Записването на говорима реч стартира при натискане върху бутона „Начало на записа“, при което се генерира автоматично уникално име на файла (табл. 3.3), който след приключване на записа ще се съхрани в папката waveFiles.

***Таблица 3.3**: Начин на образуване на уникално системно име на аудиофайла.*

Атрибут	Стойност
Уникално име на диктор:	gVVK3
Възраст на диктора:	6
Уникално име на колекция от думи	Gasteropod11
Дата /дд.мм.гггг/ и час /чч:мм:сс/ на провеждане на записа:	10.06.2014 г. 12:30:55 ч.
Уникално име на записания аудио файл: */Резултатът е автоматично образувано уникално име на записания файл /*	**gVVK3_6_Gasteropod11_10-06-2014_12-30-53.wav**

След натискане на бутон „Начало на запис“ се визуализира прозорецът показан на фиг. 3.39, в който се предоставя възможност за последователно преглеждане на отделните думи, налични в колекцията. Този прозорец заема целия екран и така позволява на детето да се съсредоточи само върху показаните изображения. Самият запис започва едва след натискане на бутона „Запис“ и приключва, когато се натисне върху бутона „Стоп“. Всички изображения, използвани в колекциите от думи и при провеждане на записите към настоящото изследване, са свалени от хранилища за свободно разпространение, като с това се цели да не се нарушат авторски права.



***Фигура 3.39**: Прозорец за запис на говорима реч,достъпен от прозорец „Запис на говорима реч“.*

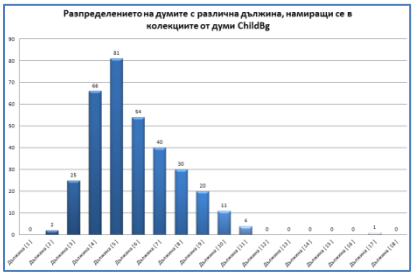
От **менюто „Помощ“** на основния прозорец може да се стартира уеб базираната помощна система към приложението, която е достъпна от бутона „Към помощта“ (фиг. 3.5). С цел улеснение на потребителите, тя е организирана по същия начин, както и самото приложение.

3.5 Анализ на думите в речника

За улеснение както на децата диктори, така и на експерта, провеждащ записите, те трябва да бъдат осъществени по увлекателен начин, така че да не се загуби интересът на детето и то да не се умори. Ето защо е необходимо записите да бъдат проведени върху отделни групи от думи. Тук тези групи от думи ще бъдат наречени **колекции от думи**. Колекцията от думи е съвкупност от няколко думи, които са тематично свързани. Пример за колекции от думи са „Диви животни“, „Домашни животни“, „Бебета на животни“. Подробно представяне на всички колекции от думи, както и принадлежащите към тях думи, може да се види в Приложение 4: Колекции от думи.

За да се провокира интересът на детето и неговото спонтанно изказване, всяка от думите е придружена от изображения и аудиофайл. Средният брой на думите в една колекция е 10, като минималният е 2, а максималният е 32. Общият брой на всички думи в колекциите е 334, а общият брой на всички думи в речника е 2457. Предвидени са две допълнителни колекции, наречени „Свободен текст“ и „Представяне“. Те не съдържат в себе си думи, тъй като се очаква да се отнасят за записи, свързани със свободен монолог или разговор на диктора с водещия експерт. Съществува още една колекция, в която няма добавени думи. Нейното име е „Неречеви звуци“ и е свързана с онези звуци, които обикновено могат да придружат едно изказване, но нямат езиково значение. Такива звуци са кихане, кашляне, поемане на въздух, издишане и т.н. Тази колекция е добавена за пълнота и предоставя възможност за изследване на неречевите звуци, както и възможност за използването им при методи за тяхното отстраняване.

На фигура 3.42 е представено разпределението на думите, които са използвани за създаване на колекциите от думи, съобразно тяхната дължина. Информацията за това е представена в табличен вид в Приложение 5. Както може да се види, най-голямо процентно съотношение заемат думите с дължина 5.



***Фигура 3.42**: Диаграма за разпределение на използваните думи за запис в колекциите от думи, спрямо тяхната дължина.*

***Фигура 3.43**: Вероятност за поява на дадена буква в думите, участвали в колекциите от думи в ChildBG.*

Под **дължина на дума** се има предвид броят на буквения състав на думата. Пресметнато е математическото очакване EX за дължина на дума в една колекция от думи. Използвана е формулата за пресмятане на математическо очакване на дискретна случайна величина X:

$$EX = \sum_{k} x_k P(x_k),$$

където k = 1, …, 18 означава броя на всички очаквани дължини на думи във всички колекции, x_k е дължината на дума с размер k, а $P(x_k)$ е вероятността за поява на дума с

дължина k. Полученият резултат за математическото очакване на дължината на една дума в която и да е колекция от думи е **5,82**.

За целите на разпознаването на реч е пресметната и вероятността за поява на всяка буква от българския език в използваните думи в колекциите от думи. Подробна информация за този анализ е представена в Приложение 6, като част от получените резултати са представени на фигура 3.43.

3.6 Изводи

В тази глава бяха изпълнени няколко задачи.

- Първо, беше изследван и представен процесът по разработването на корпус от говорима детска реч.
- Второ, беше анализиран проблемът свързан със събирането на говорима детска реч.
- И трето беше предложен и представен софтуерен продукт за събиране, обработване, съхраняване и анализ на записи на говорима реч, обособени в корпус от детска говорима реч ChildBg на диктори между 4-6 год. Беше направена обосновка за подбора на думи в речника, който ще бъде използван за направата на записите. Беше изчислено математическото очакване за появата на всяка буква в избрания речник.

Най-голямото предимство на модела е тясно специализираната насоченост и всеобхватното мултимедийно представяне на отделните думи в речника. По този начин е възможно записването на говорима детска реч с минимално участие на ръководителя на експеримента.

Като заключение може да се каже, че предложеният корпус от говорима детска реч ChildBG, притежава следните отличаващи го характеристики:

1) Управлява се от интерактивна мултимедийна система с дружелюбен графичен потребителски интерфейс;
2) Има възможност за съхраняване на богат набор от характеристики, свързани с условията на провеждане на записа, емоционалното и физическо състояние на диктора, семейната му среда и др;
3) Предоставя възможност за допълване на речника с нови думи;
4) Възможност за асоцииране на всяка дума с мултимедийни обекти;
5) Софтуерният продукт позволява пораждане на уникални абревиатури за дикторите и колекциите от думи;
6) Имената на аудиофайловете с говорима реч са уникални;
7) Има възможност за записване в корпуса на неръководена реч, чрез използването на мултимедийната колекция, придружаваща речника от думи.

ГЛАВА 4 АКУСТИЧНО И ФОНЕТИЧНО МОДЕЛИРАНЕ

4.1 Фонетичен модел и автоматично транскрибиране на български език

Според [139] звуковият състав на езика може да се разгледа в три насоки:

- Учленителна: От своя страна бива артикулационна и физиологична. Тук става въпрос за участието на артикулационните органи (говорния апарат на човек) при

продуцирането на реч. От артикулационна гледна точка звуковият състав на езика се дели на гласни и съгласни.

- *Физична:* акустична, слухова;
- *Лингвистична:* езикова, функционална, фонологична, социална.

При акустичното описание на езика се има за цел да се опишат звукови вълни, получени в резултат от работата на артикулационния апарат. Звуковете на речта представляват вълнообразни движения на въздушния поток, който идва от белите дробове, поема вибрациите на гласните струни (гласилките) и измененията получени от дейността на езика, зъбите и небцето. Както бе показано, звуковата вълна може да се опише с характеристиките: сила на звука, честота на звука, дължина и тембър.

4.1.1 Фонетична транскрипция

За да се представят точно звуковете и артикулацията на говоримата реч, се използва еднозначен начин на записване с помощта на фонетична транскрипция. Фонетичната транскрипция се различава от стандартната писменост. При нея се цели строгото съответствие между буква-фонема. Основният й принцип е, че всяка буква трябва да означава само един звук, а всеки отделен звук може да се представя с един конкретен знак (фонетичен символ).

Понастоящем съществуват няколко фонетични азбуки с различни практически приложения. Пример за компютърно четими фонетични азбуки са SAMPA [129] и нейната обновена и разширена версия XSAMPA. Друга широко разпространена фонетична азбука, използвана по-често от лингвистите, отколкото в компютърната обработка на естествени езици, е IPA (International Phonetic Alphabet) [62; 63], като в българската литература се използва термина МФА (Международна фонетична азбука).

Според [83] контекстно зависимата граматика е идеалния вариант за представяне на зависимостите между фонологичните явления. Има се предвид, че продуцирането на една фонема зависи от заобикалящите я съседни фонеми и тя може да се представи чрез правило от вида:

$$\alpha A \beta \rightarrow \alpha \gamma \beta, \tag{4.3}$$

където нетерминалният символ A е зависещата фонема, която ще се превърне в една от всичките възможни фони, в зависимост от вида на съседните й фонеми α и β.

Друг начин за представяне на контекстно зависима граматика е, чрез използването на Марковска верига. Подредбата на фонемите в думата се определя от вероятностната функция $P(W_n|W_{n-1}, W_{n-2})$, еквивалентно на продукционните правила $A_1 \overset{p_1}{\rightarrow} W_{n-2}A_2$, $A_2 \overset{p_2}{\rightarrow} W_{n-1}A_3$ и $A_3 \overset{p_3}{\rightarrow} W_nA_4$.

В наши дни представянето на синтаксиса на естествен език става с помощта на n-верига на Марков [7; 26]. Това означава, че словоредът се ръководи от n-грамната (n-gram) вероятност $p(w_n|w_{n-1}, w_{n-2}, \dots, w_1)$. Това е вероятността, че думата w_n ще бъде предхождана от поредицата от думи $w_{n-1}, w_{n-2}, \dots, w_1$. N-грамните модели се пресмятат при голям корпус от текстове, съдържащи повече от 10^9 думи. Дори при тази голяма база от данни е трудно да се получи добра статистическа информация за n>3. В някои случаи се правят правилни изчисления за n=5 или n=6, но в общия случай се използват алгоритми от динамичното оптимиране [57].

4.1.3 Фонетичен модел за транскрибиране и транслитерация на български език

При разработване на модел за формалното представяне на фонетиката на един език от

съществено значение е позицията на ударението, т.е. местоположението му в думата върху съответния звук, както и неговите съседи. За целта ще бъде създаден контекстно зависим фонетичен модел, базиран на знания. Всички правила са описани под формата на дървета с корен входната буква f_i, възли - формални правила, изградени от терминални и нетерминални символи, с които да се провери очакваното влияние на съседните звукове, и листа - очакваната фонема φ_i.

Предложения фонетичен модел е базиран на трифонния модел, който широко се използва в акустичното моделиране. При този тип модели се вземат под внимание съседните фонеми на φ_i, съответно от ляво φ_{i-1} и от дясно φ_{i+1}. Записът, който се използва за означаване на това, е

$$\varphi_{i-1} + \varphi_i - \varphi_{i+1} \quad (4.4)$$

Характерното за този модел е, че за извършване на преход от φ_{i-1} към φ_i и от φ_i към φ_{i+1} се търси най-малката цена, изчислена като вероятност на прехода.

Това, което е предложено от автора е проверяване на възможността за преход, чрез отделни правила, при които се извършва търсене в дълбочина в дървета от правила за всяка буква, до намирането на удовлетворяващо условие. След това се пресмята цената за прехода и съобразно нея се взема решение за представяне на буквата f_i като фонема φ_i.
Трябва да се отбележи, че всички използвани правила за преход са основани на сериозни проучвания на фонетиката на българския книжовен език, част от тях представени формално в Глава 2. Тази методика е позната под наименованието методи, базирани на знания (knowledge-based method). Това е традиционен метод за намиране на фонеми с най-голямо съответствие [87].

За да се определи вероятността за грешка на модела, е необходимо да се изчислят сходствата между отделните фонеми. За целта се използва алгоритъм, състоящ се от две стъпки. На входа се подава текст, написан на български, а за изходен език са използвани символите от международната фонетична азбука (IPA).
В резултат от направеното изследване на фонологичните особености на фонетичната система на българския език във втора глава, бе построено йерархично-фонетично дърво (фиг. 4.4), което еднозначно определя звуковете в българския език. Въз основа на това дърво и правилата за дистрибуция и съчетаемост на фонемите, е изграден фонетичен модел, за автоматично транскрибиране на ортографичен (написан) текст. Всички звукове в дървото се намират на отделни нива, определящи дълбочината на съответния възел, и всяка дъга има тегло, отразяващо вероятността за преминаване към следващото подниво.

За да се определи сходството между отделните фонеми, ще се търси разстоянието, на което речевите звуци отстоят един от друг в йерархично-фонетичното дърво (фиг. 4.4)

Използваният алгоритъм е базиран на метод в [75], като в голяма степен е изменен и адаптиран за българския език от автора на настоящата дисертация. Състои се от две стъпки: на първата стъпка се извършва търсене отгоре-надолу чрез йерархично-фонетичното дърво (което по своята същност е насочен граф) и на втората стъпка се пресмята разстоянието (пътя) отдолу-нагоре между намерените фонеми. Той е базиран.

Стъпка 1: Търсене отгоре-надолу

На фигура 4.4 е представено йерархично-фонетично дърво на основните фонеми в българския книжовен език. Всички фонеми са групирани в отделни класове (групи, клъстери) в зависимост от начина на учленяване, съобразно звученето им и техните специфични особености. Класовете в дървото са представени като възли и на всяко

следващо ниво са разпределени в подкласове. Листата на дървото са фонемите според IPA, а непосредствено преди тях (предпоследното ниво) са техните аналози спрямо българската фонетична транскрипция. Това е направено за пълнота на изложението.
Нека k е броят на нивата и L_i, $i = 1, \dots, k-1$, да бъде потребителски дефинирана стойност на сходство за i-тото ниво. Ще оценим k и L_i, така че да се изведе възможно най-бързо (за най-малък брой стъпки) точната транскрипция чрез фонетичния модел, който е предложен.

На тази стъпка е необходимо да се намери към кои от класовете принадлежат съпоставените фонеми, и съответно през колко нива трябва да се премине, за да се стигне до тях. L_i в този случай ще бъде номерът на общото родителско ниво и за двете търсени фонеми. Колкото е по-голямо това число, т.е. имат по-горен родителски възел, толкова те се намират на по-голямо разстояние по между си и съответно има по-малка вероятност при артикулация и при разпознаване, да бъдат взаимно заменени.

Йерархично-фонетичното дърво е разделено на поднива, номерирани от 0 (корена) до 9 (листата) и в тях са включени следните класове (възли):

- *Ниво 0:* „Звуковете в българския език“;
- *Ниво 1:* „Съгласни“ и „Гласни“;
- *Ниво 2:* За съгласните имаме „Висока тоналност“, „Средна тоналност“ и „Ниска тоналност“, а при гласните „Предни“ и „Задни“;
- *Ниво 3:* За съгласните имаме „Алвеолни“, „Предннебни“, „Веларни“, „Лабиални“, а за гласните – „Нелабилани“ и „Лабиални“;
- *Ниво 4:* За съгласните – „Алвеодентални“, „Билабиални“, „Лабиодентални“, а за гласните „Висока тоналност“, „Средна тоналност“ и „Ниска тоналност“;
- *Ниво 5:* Само за съгласните – „Съскави“, „Плавни“, „Шушкави“;
- *Ниво 6:* Само за съгласните – „Преградни“, „Проходни“, „Преградно-проходни“, „Литерални“, „Вибранти“, „Назални“, „Глайд“;
- *Ниво 7:* За съгласните - „Звучни“ и „Беззвучни“, а за гласните – „Широки“ и „Тесни“;
- *Ниво 8:* Всички звукове съгласно българската фонетична транскрипция;
- *Ниво 9:* Всички звукове съгласно международната фонетична азбука (IPA).

Това дърво е от голямо значение при построяване на формалните правила, тъй като видът на класа, към който спада даден звук, определя начина на учленяване и влиянието му върху неговите съседи. При определяне на неформалните символи се използват главни латински букви, като съгласните се означават с C, а гласните – с V. Вида на класа, към който спадат съгласните, се означава с горен индекс. Пример за това са класовете от всички вибранти и от всички преградни, като членовете на са самите фонеми според IPA:

$$C^{\text{вибранти}} = \{r, r^j\} \tag{4.5}$$

$$C^{\text{преградни}} = \{d, d^j, t, t^j, g, g^j, k, k^j, b, b^j, p, p^j\} \tag{4.6}$$

Когато означаваме само отделен елемент от този клас тогава използваме и долен индекс за определяне на неговата позиция в класа, като за първа позиция винаги използваме 0. Например звука t се означава като $C_2^{\text{преградни}}$, а i-тия звук от даден клас от съгласни като $C_i^{\text{клас}}$.

Стъпка 2: Оценяване отдолу-нагоре

В първата стъпка бяха намерени двете търсени фонеми φ_s и φ_t. Сега е необходимо да се определи разстоянието помежду им, което всъщност определя тяхното сходство.

Сходството между φ_s и φ_t се получава чрез стойността за сходство на *i*-тото ниво, което е родителско за двете фонеми, т.е.

$$d(\varphi_s, \varphi_t) = L_i \tag{4.7}$$

Например родителският възел на фонемите [d] и [t] се нарича „Преградни“ и е на ниво 6, т.е. сходството между тях е:

$$d([d], [t]) = L_6 \tag{4.8}$$

На всяка дъга са поставени теглови коефициенти, които са пресметнати, така че сумата на излизащите дъги от един възел да е равна на 1, а тяхната стойност е равна на вероятността за поява на едното или другото подниво.

Коефициентът на сходство се пресмята чрез цената на прехода от едната фонема към другата, като сумата на всички дъги между двете фонеми (които всъщност са листата в дървото). Това се представя като:

$$r(\varphi_s, \varphi_t) = \sum_{m_s} p_{m_s}(l, k) + \sum_{m_t} p_{m_t}(q, z) \tag{4.9}$$

където m_s е броят на всички дъги от листа, съдържащ фонемата φ_s, до общия възел от ниво i, а m_t аналогично е броят на всички дъги от фонемата φ_t до възела от ниво i. С $p_{m_s}(l, k)$ е означена вероятността за прехода между два съседни върха *l* и *k*.

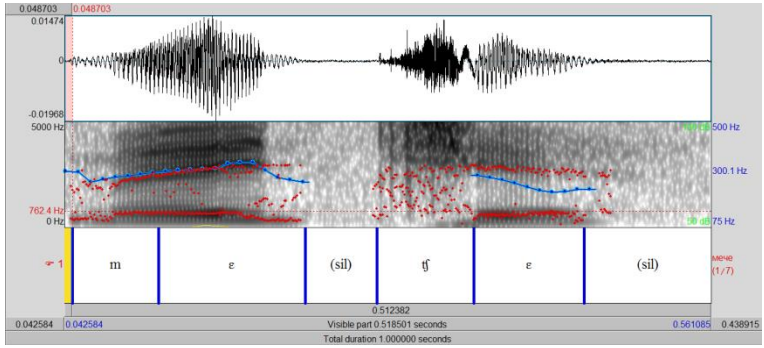
Цената за прехода в разглеждания по-горе пример е

$$\begin{aligned} r([д], [т]) = {} & p([д], [д]) + p([д], \text{Звучни}) + p(\text{Преградни}, \text{Беззвучни}) \\ & + p(\text{Беззвучни}, [т]) + p([т], [т]) \\ = {} & 1 + 0.5 + 0.5 + 0.5 + 0.5 + 1 = 4 \end{aligned} \tag{4.10}$$

Колкото е по-малка цената за прехода между две фонеми, толкова вероятността за настъпване на грешка при акустичното моделиране на говорима реч е по-голяма, тъй като двата звука имат по-сходни свойства.

С помощта на предложения по-горе алгоритъм лесно може да се определи сходството между отделните фонеми и коефициентите на сходство.

Сега трябва да се определят формалните фонетични правила за преобразуване на една буква (букви) във фонема. Най-голямото предизвикателство при изграждане на тези правила е ударението в българския език. Обикновено акустично ударението се характеризира с по-висок интензитет, по-голяма продължителност и по-висока честота на основния тон.



***Фигура 4.5**: Изказване на думата “МЕЧЕ” на диктор на 6 години.*

Според [139] ударението в българския език е силово (динамично), тъй като при

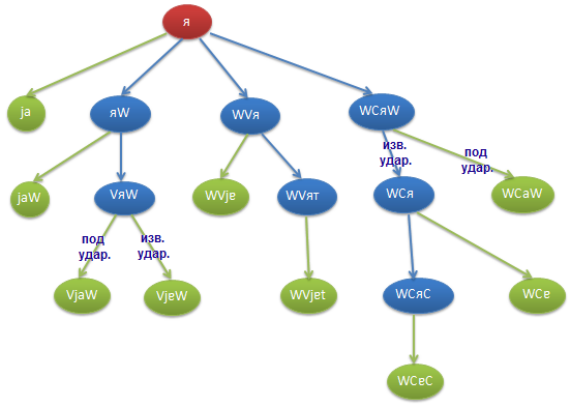
ударената сричка се наблюдава ясна и отчетлива артикулация на всички нейни звукове. Освен това ударението в българския език притежава свойството музикалност, тъй като ударената гласна се произнася с по-висок тон от неударената. Пример за това е изказване на диктор на 6 години (фиг. 4.5), при който средният период на основния тон на звука є под ударение е 303 Hz, докато на същия звук в края на думата е 239 Hz.

Ето защо при изграждане на фонетичния модел от съществено значение е използването на ограничения, свързани с позицията на ударението. В представените формални правила за означаване на ударението ще бъде използван специалния символ ` (апостроф), тъй като програмно е по-трудно отразяването на ударението по друг начин. Например ударението на буквата [à] ще изглежда като [`a].

При построяване на формалните правила са използвани въведените по-горе означения, а именно:

- W – всяка последователност от звукове;
- C – съгласен звук, един или няколко;
- V – гласен звук, един или няколко;
- с малки букви са означени терминалните символи, които са или част от буквите в българската азбука, или от фонетичната азбука.

За всяка от буквите в българския език е построено дърво, отразяващо нейното звучене при различните условия, съобразно нейните съседи. На фигура 4.6 е представено като пример фонетичното дърво от правила за буквата Я, която е дифтонг (състои се от два звука). За всички останали букви се използва същият принцип.



***Фигура 4.6**: Формално дърво от правила за представяне на буквата „Я“.*

На входа се подава последователност от букви $W = f_1 f_2 f_3 f_4$ … Част от тези букви могат да бъдат интервали (паузи), препинателни знаци, цифри или букви от други азбуки. Изследваният масив се състои от буквите на българската азбука, а изходния масив е фонетичните символи спрямо IPA. Стъпките, които се изпълняват при този алгоритъм, са следните:

Стъпка 1: Извлича се първият символ f_1 и се проверява, дали е елемент от изследвания масив.

Стъпка 2: Ако това е така, се проверява кой елемент е от масива, запазва се и се преминава към неговия съсед f_2, за да се определи какво е състоянието на f_1. Това се

осъществява чрез търсене в дървото от правила за изследвания звук f_1. Ако не е, се преминава към следващия елемент.

Стъпка 3: Ако е необходимо, се проверява и следващият символ f_3, при което алгоритъма се връща отново към корена f_1 и се взема решение за ноговото фонетично представяне като φ_i.

Стъпка 4: Извлича се следващият символ по ред, който ще бъде означен по-общо с f_i и се проверява дали е част от изследвания масив.

Стъпка 5: Ако е така се преминава към стъпка 6, в противен случай се повтаря стъпка 4, докато не се стигне до края на входния последователност от символи. Ако няма повече символи, се преминава към стъпка 8.

Стъпка 6: Извличат се съседите на f_i, т.е. f_{i-1} и f_{i+1}, и се проверява, какво е отношението между тях съгласно дървото от правила за f_i. При необходимост се проверяват съседите f_{i-2} и f_{i+2}.

Стъпка 7: Преминава се към стъпка 4.

Стъпка 8: Край на алгоритъма.

4.1.4 Програмна реализация за тестване на предложения фонетичен модел

Въз основа на предложения фонетичен модел е създадена уеб базирана система за автоматично фонетично представяне на текст на български език, наречен BulPhonetic (фиг. 4.8 и фиг. 4.9).

***Фигура 4.9**: Изглед на уебсайта BulPhonetic (страница Text to Phonetics).*

Към настоящия момент системата работи само с фонетичната азбука SAMPA и извършва транслитерация на латиница. Основният работен прозорец е с интерфейс изцяло на английски език, с цел по-лесното му използване от чужденци, които използват или изучават български език. Системата е създадена с Visual Studio 2013 и езика C#.

Източник [142] бе използван за изграждане на алгоритъма за автоматична транслитерация на собствени имена на български език и представянето им чрез латиница. На този етап от разработка главните букви се преобразуват в малки. Функцията е достъпна на електронния адрес, посочен в [70].

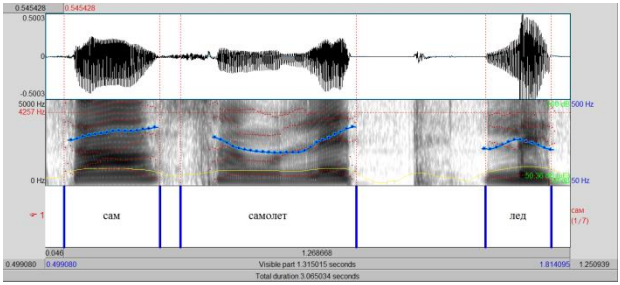
4.3 Кепстрален анализ на речевия сигнал

За реализиране на правилното обучение на системата за разпознаване, първо е необходимо записаните данни да бъдат етикетирани (mapping) и впоследствие прегледани за верифициране на получените данни. Но етикетирането на непълни и неясни данни също се оказва сложен процес.

Като пример може да се използва изказването: „САМ САМОЛЕТ ЛЕД“. Дикторът е момиче на 5 години и 6 месеца, без говорно-комуникативни нарушения. Записът е направен в тиха, нешумоизолирана стая. Сегментирането на думите е направено ръчно с програмата Praat.

От фиг. 4.11 се вижда, че двете думи „САМ“ и „САМОЛЕТ“ трудно се разграничават, а

между думите „САМОЛЕТ“ и „ЛЕД“ има шум в резултат от вдишването и издишването на въздуха от детето. При стандартно определени параметри за изчистване на шума на реч, продуцирана от възрастен диктор, буквата „С“ от думата „САМОЛЕТ“ остава извън диапазона както на формантите, така и на периода на основния тон. Ако това се подаде към система за разпознаване, то думата, която следва да се получи на изхода, ще бъде „АМУЛЕТ“, което е съвсем различна дума от първоначалната. Ето защо често се прибягва до нормализация на спонтанната детска реч.



***Фигура 4.11**: Анализ на изказването „САМ САМОЛЕТ ЛЕД“, на диктор момиче на 5 год и 6 мес.*

Формантите възникват и се виждат на спектрограмата около честоти, които съответстват на резонансите на говорния тракт. Но има разлика между оралните (устенните) гласни от една страна, и съгласните и назалните гласни от друга. За съгласните има и антирезонанс в говорния тракт, в една или в повече честоти, поради стеснение на устните. Антирезонансът е обратното явление на резонанса, така че съпротивлението е относително високо, отколкото ниско. Следователно резонансът и антирезонансът намаляват или елиминират формантите в близост до тези честоти, така че формантите на тези звуци са отслабени или направо липсват, при детайлно разглеждане на спектрограмата. Ето защо е трудно да се видят форманти по-долу от 3000 Hz за двете копия на [s] в спектрограмата по-горе. В случай на безгласната проходна [s], както е при „САМ“ и „САМОЛЕТ“, има апериодично съскане поради принудителна струя въздух, удряща се в предните зъби, което поражда резонанс и води до смущение на всички честоти. Смущението се разпространява чрез говорния тракт, и отново преминава свободно през честоти близки до резонансната, но не преминава свободно в честотите между резонансите, създавайки силни пикови форманти с по-слаби склонове между тях.

За справяне с този проблем тук е изследвана нормализацията на акустичните характеристики. Тя се извършва, чрез нормализиране на кепстралните изменения (cepstral variance normalization) [45]:

$$\hat{c}_i(t) = \frac{c_i(t) - \mu_i}{\sigma_i} \qquad i = 1, \dots, N, \tag{4.11}$$

където i е кепстралния индекс, t е анализираният кадър от време, N е броят на кепстралните коефициенти, μ_i е средната стойност, а σ_i е стандартното отклонение.

Средната стойност на кепстралните коефициенти за един кадър (фрейм) се изчислява, чрез формулата:

$$\mu_i = \frac{\sum_{i=0}^{L} c_i(t)}{L}, \tag{4.12}$$

където L е броя на кепралните коефициенти.

Стандартното отклонение за кепстралните коефициенти се изчислява, чрез:

$$\sigma_i = \sqrt{\frac{\sum_{i=0}^{L}(c_i(t)-\mu_i)^2}{L-1}}\ , \qquad (4.13)$$

За извличане на мел-честотните кепстрални коефициенти на думата „САМ“ от горния пример се използват следните параметри:

- брой на коефициентите: 12;
- ширина на прозореца: 0,015 sec;
- времева стъпка: 0,005 sec;
- позиция на първия филтър по мел-скалата: 100 mel;
- разстояние между филтрите (mel): 100,0 mel;

Извлечени са 37 коефициента с максимална честота 3900 Hz, а броят на получените кадри е 51. Полученият резултат за първите 8 кадъра е показан на таблица 4.2.

***Таблица 4.2**: Таблица с кепстралните коефициенти от първите 8 кадъра, при анализ на думата „САМ“ на 5 годишен диктор.*

Коеф.	кадър [1]	кадър [2]	кадър [3]	кадър [4]	кадър [5]	кадър [6]	кадър [7]	кадър [8]
c [0]	658,6161003	711,7083857	707,301714	745,7056874	769,9061848	760,3184429	796,8999612	854,7503358
c [1]	-82,68973228	-80,07029384	-123,5399376	-128,9533492	-99,04191301	-106,6179632	-103,9110576	-97,33075196
c [2]	111,9320087	98,15160957	83,62583157	78,09599509	117,7653784	105,695689	116,9150856	104,4252773
c [3]	124,4615106	130,7559402	123,3159147	108,4860676	113,3258741	103,6669711	108,0181332	106,9935875
c [4]	-25,08128192	-23,60262219	-32,06339526	-43,46407943	-31,4854591	14,73878057	-26,39030762	-64,28594318
c [5]	64,33025032	62,27630028	69,96374732	57,9909979	62,27131299	78,42010183	64,51030016	39,09616165
c [6]	-42,95578169	-24,57917258	1,316159861	-51,13517469	-26,31671847	-50,5716367	-40,57169523	-15,77408201
c [7]	33,27055484	27,44929832	23,47603204	4,21599981	17,50305738	22,6704666	20,34414832	28,5050777
c [8]	-9,358029448	-9,856245892	-6,217155461	-7,687636099	-11,70262464	-2,171225273	9,553179011	-4,241652543
c [9]	46,78866769	32,73977396	53,84356139	32,18949372	21,937853	28,7426128	31,50172895	6,157359337
c [10]	-17,70564698	-8,253420045	-4,71960493	28,54417979	-0,952015053	-23,13500917	-21,74465355	-13,38079434
c [11]	37,14258339	27,452911	32,24717732	45,40928165	29,2962091	19,85261598	19,59400932	35,71901642
c [12]	-10,99725506	-19,52102457	-10,82153027	-18,38735657	-15,75419542	-23,26663757	-21,15615309	-9,091208772

След прилагане на нормализацията на кепстралните коефициенти се получи резултатът представен на таблица 4.3.

***Таблица 4.3**: Таблица с нормализирани кепстрални коефициенти от първите 8 кадъра, на кепстралните коефициенти от таблица 4.2.*

Коеф.	Кадър [1]	Кадър [2]	Кадър [3]	Кадър [4]	Кадър [5]	Кадър [6]	Кадър [7]	Кадър [8]
$\hat{c}$ [0]	0,917990096	0,929979636	0,921550369	0,925993298	0,93057673	0,930140786	0,931865446	0,937207723
$\hat{c}$ [1]	-0,234779515	-0,219504416	-0,280984195	-0,264651756	-0,229439335	-0,240369037	-0,228297415	-0,206739834
$\hat{c}$ [2]	0,067867542	0,039233588	0,018861145	0,017197819	0,059991002	0,046290069	0,056106621	0,035674724
$\hat{c}$ [3]	0,087351577	0,086567722	0,076307347	0,058566841	0,054064416	0,043550959	0,044648154	0,038760609
$\hat{c}$ [4]	-0,145195341	-0,137526094	-0,148583871	-0,148277979	-0,139253774	-0,076517041	-0,128457723	-0,167035729
$\hat{c}$ [5]	-0,006155699	-0,012849269	-0,000912934	-0,010170464	-0,014091682	0,009463444	-0,011385987	-0,042819727
$\hat{c}$ [0]	-0,172991129	-0,138943825	-0,100271333	-0,158720392	-0,132353681	-0,164697078	-0,146722063	-0,1087476
$\hat{c}$ [1]	-0,054455161	-0,06341022	-0,068197818	-0,083372629	-0,073855776	-0,065807932	-0,068267984	-0,05554516
$\hat{c}$ [0]	-0,120744855	-0,117569456	-0,111174821	-0,099576663	-0,112844365	-0,09934839	-0,082165773	-0,094891118
$\hat{c}$ [1]	-0,033433784	-0,055729643	-0,024244793	-0,045293217	-0,067935475	-0,057609514	-0,053898032	-0,082396463
$\hat{c}$ [0]	-0,13372584	-0,115242514	-0,109007313	-0,050255465	-0,098492668	-0,127653021	-0,122474542	-0,105872009
$\hat{c}$ [1]	-0,048433953	-0,063404975	-0,055502733	-0,027297547	-0,05811232	-0,069612504	-0,069234095	-0,046877445
$\hat{c}$ [0]	-0,123293937	-0,131600535	-0,117839051	-0,114141847	-0,118253073	-0,127830741	-0,121716607	-0,100717972

Целта на нормализирането на измененията на акустичните характеристики е да се намалят различията между речта на отделните диктори и да се ограничи ефектът на коартикулацията. Най-сериозното ограничение на този метод е големият обем от данни, които трябва да се обработят.

4.4 Използване на интерактивен самоорганизиращ се метод за анализ на данни при класификацията на акустичните характеристики

Предимствата на алгоритъма ИСОМАД (Интерактивен СамоОрганизиращ Метод за Анализ на Данни) се дължат на способността му да отстранява клъстерите с малък брой елементи, да разцепва онези от тях, които имат несъвместими свойства и да обединява други, които притежават сходни такива. Изходните резултати обаче зависят изцяло от евристичните входни параметри *N*, *K*, *I*, *L*, θ_N, θ_S, θ_C. Тези параметри са [83]:

- *N* – параметър, определящ броя на входните образи;
- *K* – параметър, определящ очаквания брой клъстери;
- *I* – параметър, определящ максималния брой итерации;
- *L* – параметър, определящ максималния брой двойки центрове, които могат да бъдат обединени (lumping)
- θ_N – параметър, определящ минималния брой елементи влизащи във всеки клъстер (използва се за ликвидиране на ненужните клъстери);
- θ_S – параметър, определящ максималното стандартно отклонение за всеки клъстер (използва се в операцията разцепване);
- θ_C – параметър, определящ минималното разстояние между два центъра на клъстери, т.е. това е параметър характеризиращ компактността (използва се в операцията сливане).

В резултат на много тестове бе достигнато до извода, че разпределянето на образите по клъстери при класическия алгоритъм ИСОМАД, може да бъде подобрено. След внимателно изследване на алгоритъма бе установено, че в стъпките 7 и 10 се съдържат еднакви условия за проверка, които се отнасят до това дали текущият брой на клъстерите (NC), е два пъти по-малък от очаквания, т.е $Nc \le K / 2$.

Целият алгоритъм е представен в Глава 2 и подробно описан в [23; 83], тук ще бъдат представени само двете стъпки, в които са променени условията:

Стъпка 7: Ако текущата итерация е последна, то параметърът определящ минималното разстояние между два клъстера приема стойност $\theta_C = 0$ и се преминава към *Стъпка 11*.

Ако текущият брой на клъстерите е по-малък от половината от очаквания брой, т.е. $N_C \le K / 2$ (много малко клъстери), се преминава към *Стъпка 8.*

Ако текущата итерация има четен пореден номер или броят на получените клъстери е по-голям от удвоената стойност на очаквания брой, т.е. $N_C > 2K$ (твърде много клъстери), се преминава към *Стъпка 11*.

В противен случай изпълнението на алгоритъма се преустановява.

Стъпка 10: Ако за някое максимално средно-квадратично отклонение $\sigma_{j_{max}} (j = 1, \dots, N_c)$ на j-тия клъстер са изпълнени условията $\sigma_{j_{max}} > \theta_S$, т.е. максималното средно-квадратично отклонение е строго по-голяма от максималното стандартно отклонение за j-тия клъстер, и:

а) средната стойност на разстоянията $\overline{D}_j$ между обектите е строго по-голямоа от обобщеното средно разстояние $\overline{D}$, т.е. $\overline{D}_j > \overline{D}$, и текущият брой на обектите, попадащи в j-тия клъстер N_j е строго по-голям от $2(\theta_N + 1)$, където θ_N е минималния брой образи, които могат да принадлежат на един клъстер, т.е. $N_j > 2(\theta_N + 1)$; или

б) полученият общ брой на клъстерите N_C е по-малък от $N_C \le K/2$, то клъстерът с център z_j се разцепва на два нови клъстера с центрове z_j^+ и z_j^-, като съответно се добавя и

изважда зададена константна величина γ_j към максималната компонента на вектора $\sigma_{j_{max}}$. За стойност на γ_j може да се вземе максималната средно квадратична компонента, т.е. $\gamma_j = k\sigma_{j_{max}}, 0 < k < 1$.

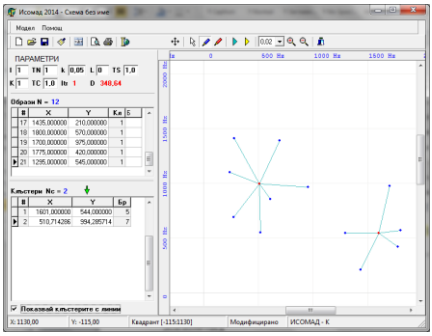
Препоръчителна стойност за k е *0,5*, тъй като стойността на γ_j трябва да бъде достатъчно голяма, за да се отличават разликите в разстоянията между образа и двата нови центъра, и същевременно да бъде достатъчно малка, за да не настъпят промени в общата структура на процеса на клъстеризация. След това се изтрива старият център z_j и N_C=N_C+1.

Ако разцепването е извършено на тази стъпка, се преминава към *Стъпка 2*.

В противен случай се изпълнява следващата стъпка.

Стъпка 7 е стъпка на разклонение в алгоритъма, при която има три възможности: едната е да се премине към разцепване на клъстери; другата – към сливането им; третата – към безусловен изход от алгоритъма.

Модификацията на алгоритъма се състои в промяната на условията в *Стъпка 7* и в *Стъпка 10*. На местата в алгоритъма, в които се проверява дали текущият брой на клъстерите N_c е достигнал половината от очаквания резултат, т.е. $N_c \leq K/2$, това условие да бъде заменено с условието: $N_c \leq K$.



Фигура 4.13: *Изглед на интерфейса на програмния продукт „ИСОМАД 2014“*

С тази промяна се запазва цялостната концепция на алгоритъма, т.е. получената модификация може да се прилага в процеса на класификация. С тази промяна се цели да се постигне по-детайлно разпределяне на образите с общи свойства по класове, което при първоначалния вариант на алгоритъма се достига трудно, тъй като ИСОМАД зависи изцяло от входните параметри, които имат евристичен характер. Не винаги е лесно определянето на броя на класовете (клъстерите), на които може да се разделят образите.

От ограничителните условия се вижда, че преди да бъдат променени, броят на очакваните клъстери няма да бъде достигнат никога, защото алгоритмът ще спре своето изпълнение след като са сформирани половината клъстери. В този случай, за да бъдат установени класовете на разпределение на образите, експертът ще трябва да въведе два пъти повече клъстери от този, които реално се очаква да има.

Предложената модификация е имплементирана в програмния продукт „ИСОМАД 2014“, представен на фигура 4.13. Този софтуер ще бъде използван за акустичното моделиране на данните, получени след анализиране на записите от говорима детска реч.

4.5 Изводи

В тази глава бяха постигнати следните резултати:

- Беше представено йерархично-фонетично дърво за еднозначно определяне на звукове в българския език.
- Беше описан контекстно-зависим фонетичен модел за определяне на позицията на всеки от звуковете и възможността му за заместване с друг звук (съпоставяне

между звуковете).

- Беше представен алгоритъм за автоматично транскрибиране, като част от използвания фонетичен модел, основан на правила. Накратко бе представена уеб системата BulPhonetic, в която е имплементиран представения алгоритъм.
- Беше направен кепстрален анализ на детска реч и реализирана нормализация на получените кепстрални коефициенти.
- Описана бе модификация на ИСОМАД, който ще бъде използван за класифициране на акустичните характеристики от аудиозаписите, събрани в корпуса от говорима детска реч ChildBG.

Глава 5. ЕКСПЕРИМЕНТАЛНИ ЗАДАЧИ

5.1 Събиране на данни от говорима детска реч

В Глава 3 беше представен интерактивен мултимедиен софтуерен продукт за създаване и поддържане на корпус от говорима детска реч ChildBG. Целевата група в настоящото изследване се състои от деца на възраст от 4 до 6 години, но за пълнота на корпуса е записана говорима реч на диктори и от други възрастови групи. Освен това в корпуса има записи на деца в разглеждания възрастов диапазон, които са от ромски етнос.

За изследване на речевата вълна и извличане на акустичните характеристики ще бъде използван програмният продукт Praat, разработен от университета в Амстердам и предоставящ множество възможности за цифровата обработка на аудиосигнали [13].

5.1.1 Методи на събиране на данни от говорима детска реч

В този раздел ще бъде представено сравнение на резултатите от стандартното събиране на говорима детска реч и събирането на реч чрез използването на софтуера за работа с корпуса ChildBG. Всички направени записи са обработени и колекционирани в ChildBG и могат да се използват за бъдещи изследвания.

Първият експеримент представлява стандартен метод за записване на детска говорима реч. Състои се в провеждането на интервю (запис) с помощта на педагог-логопед, който в игрова форма произнася желаната дума и детето повтаря след него. Записите са направени със софтуера за фонетична обработка Praat, като са интервюирани 20 диктора. По време на записа експертът произнася желаната дума и детето машинално я повтаря. За да не скучае и да не се разсейва по време на записа, на детето му се дават игрови карти. Така то се концентрира само върху подадената му от експерта дума.

Предимство на стандартния метод е, че за кратък интервал от време с един диктор са проведени повече сесии, средно около 18 на брои и средната продължителност на една от тях е около 00:05:05 ч.

Тук възниква проблема, при който децата в 75% от случаите повтарят не само думата, но и интонацията, дикцията и тона на експерта. Така се губи възможността да се уловят реалните способности на диктора за продуциране на реч. Освен това повечето записи са придружени с допълнителен шум, като обръщане на карти, въздишане на децата и т.н. Неудобство е, че експертът трябва да си води бележки за текущото състояние на околната среда, емоционалното и физическо състояние на диктора.

Резултатите от проведения стандартен метод са представени на първия ред в таблица 5.1.

***Таблица 5.1**: Сравнение на данните от направените записи (сесии), в зависимост от използвания софтуерен продукт.*

Използван софтуер за запис	Продължи-телност на всички записи	Брой на всички записи	Бр. на дикторите	Средна продължи-телност на запис	Осреднен брой на записите за един диктор
Praat	1:41:44	360	20	00:05:05	18,00
ChildBG	01:24:07	203	21	00:03:36	8,43
Общо	03:05:51	563	43		

Чрез използването на интерактивната мултимедийна система за работа с ChildBG, бяха интервюирани 21 диктора. Проведени са общо 152 сесии (записи), като средната продължителност на една сесия (запис) е 00:03:12 ч. Продължителността по време на една сесия е съизмерима с тази, проведена заедно със специалист-логопед и с помощта на софтуера Praat.

По време на провеждане на интервюто с предложения софтуер експертът, ръководещ записа, не подпомага диктора, а само сменя изображенията на екрана. Това позволява на диктора да произнесе неочаквана дума или думи (неръководена реч). При експериментите бяха наблюдавани случаи, при които детето диктор не разпознаваше избраните изображения за представяне на съответната дума или не знаеше какво има на тях. То назоваваше друга дума или казваше „Незнам“. Срещат се и редки положения, в които детето задава въпроси и очаква от експерта да му назове думата. Интересното при всички тези примери е, че дикторът не се старае да имитира интонацията и тона на експерта, а продуцира очакваната или неочакваната дума / думи по свой собствен начин. Това позволява получената реч, да се нарече неръководена и спонтанна.

Недостатък на този метод е, че в редки случаи дикторите не разпознават използваните изображения, и тогава експертът се намесва, като произнася съответната дума. Тук за разлика от първия описан метод се наблюдава, че детето диктор е заинтересовано от този процес, не се уморява и има склонност да разкаже повече неща за едно изображение. Освен това системата позволява да се опишат всички допълнителни шумове, неречеви звуци и затрудненията на диктора по време на речеобразуването.

Резултатите от интервютата с използването на интерактивния мултимедиен софтуер за работа с ChildBG са представени на втори ред в таблица 5.1.

Най-важните предимства и недостатъци на двата метода за събиране на говорима детска реч са обобщени в таблица 5.2.

***Таблица 5.2:** Сравнение на плюсовете и минусите на двата начина на провеждане на записите*

Използван Метод	Плюсове	Минуси
Стандартен метод	• Повече записи;	• Наличие на допълнителни шумове; • Дикторът имитира тона и интонацията на експерта.
ChildBG	• Забавно и интересно за детето диктор; • Липса на умора; • Продуциране на неръководена реч;	• По-малък брой на сесии с един диктор

5.1.2 Анализ на събраните данни от говорима детска реч

Записите се направени в различни околни среди. Част от тях са записани в университетска зала, в зала на учебното заведение, в което учи дикторът, и в домашна обстановка. Използваните стаи са отделени, но не са шумоизолирани или специално подготвени за записване на идеален звук. Това е и причината в повечето от тях да присъства страничен шум. Системата управляваща ChildBG позволява експерта да изброи

наличните допълнителни шумове към съответния запис, което от своя страна облекчава изследването на околния шум, ехото и постоянния шум.

Всички записи на говорима реч са направени с микрофон Philips SBM MDI50. Това е динамичен микрофон с честотен обхват от 85 Hz до 11 000 Hz, чувствителност на мембраната -74 dB и импеданс 600 Ohm. Използването на един и същи микрофон за всички налични данни в корпуса е добър вариант, тъй като не се разглежда като допълнителен фактор, влияещ върху качеството на получения звук. Така може да се насочат повече изследователски усилия върху влиянието на околната среда при записания звуков сигнал и характеристиките на речта на съответния диктор.

По време на интервюто част от децата говорят тихо и плахо пред микрофона. Това допълнително влияе върху качествата на продуцираната от тях реч. Повечето от децата не са повтаряли многократно, очаквания текст. При използване на изображенията онагледяващи думите в избраните колекции в ChildBG, децата диктори продуцират спонтанна реч, като назовават по различен начин съответната дума. В някои от случаите думите са продуцирани в тяхната форма, което е отразено като влияние на семейната среда и редица други фактори, представени като данни в системата. Това позволява освен реалното отразяване на околната среда и изследване на текущото емоционално състояние на дикторите, когнитивните им познания, психичното и лингвистичното им развитие.

По времето на някои от записите децата се кикотят, смеят, кашлят или подсмърчат. Това са така наречените неречеви звуци, и наличието им е отразено с помощта на една от функциите на софтуера за работа с ChildBG. Някои от тези звуци са изследвани и представени по-долу.

В записаната реч са наблюдавани интересни събития, характерни за малката възраст на децата. Такива са способността им да пропускат трудните звукове, да заменят един звук с друг, да разместват местата на звуковете или да започнат изказването с много висок тон, а да го приключат с много нисък, сравним с шепот. Всички тези любопитни способности на децата могат да се отразят в системата за управление на ChildBG, като разместването и пропускането на звукове е отразено с помощта на специални символи.

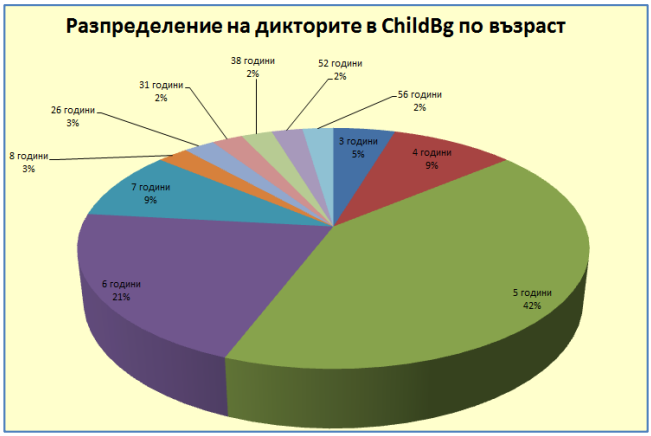
Общият брой на направените записи е 563, а на дикторите е 43. Броят на децата диктори е 38, а на възрастните е 5 (табл. 5.3).

Таблица 5 3:*Разпределение на диктори по пол и възраст.*

Възрастови категории на дикторите	Възраст на дикторите	Пол		Общ брой на диктори	Общ бр. на дикторите по възраст
		Мъже	Жени		
Деца	3 години	2	0	2	38
	4 години	2	2	4	
	5 години	10	8	18	
	6 години	2	7	9	
	7 години	2	2	4	
	8 години	0	1	1	
Възрастни	26 години	0	1	1	5
	31 години	0	1	1	
	38 години	1	0	1	
	52 години	0	1	1	
	56 години	1	0	1	
Общо по пол:		**20**	**23**		
Общо:		**43**			

За онагледяване на тези резултати е използвана кръгова диаграма, на която е отразено процентно разпределение на дикторите, според тяхната възрастова категория, участвали в

провеждането на записи в корпуса ChildBG (фиг. 5.1).



***Фигура 5.1**: Диаграма на процентното съотношение на диктори, разпределени по възраст.*

Както може да се види от диаграмата (фиг. 5.1) и от таблицата (табл. 5.3), са интервюирани диктори от 11 възрастови групи. За пълнота на получените данни, освен деца между 4 - 6 години, корпусът съдържа аудиозаписи и на деца на 3 год., 7 год., 8 год. и възрастни диктори. Изследваната целева група от деца обхваща 72 % от всички диктори, което удовлетворява условието за адекватно отразяване на тяхната комуникативна способност.

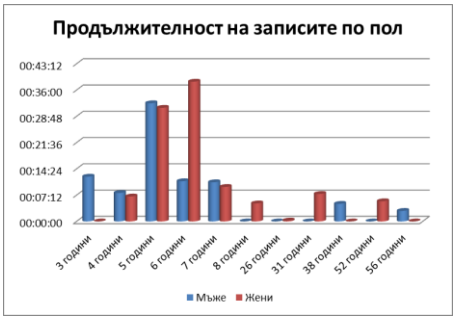
След провеждане на интервютата бе установено, че по-малките деца се затрудняват и уморяват от дългите колекции от думи, като например „Училище“ (32 думи). Софтуерът предоставя възможност за направата на подробен анализ на колекциите от думи в корпуса, и конкретно думите, при които се е наблюдавала трудност. Тези анализи не са цел на настоящата дисертация и няма да бъдат представени тук. В предходната глава бе обяснен начинът на подбор на използвания речник от думи за произнасяне и конструирането на колекции от думи, които са тематично обособени.

Общата продължителност на всички направени записи на деца е 02:44:51 часа, а на възрастни 00:21:00 часа. Най-голям дял имат записите на 5-годишните деца, следвани от тези на 6 г., 7 г. и т.н. Цялата продължителност на всички налични записи в корпуса ChildBG е 03:05:51 часа. Продължителност на записите на момчета е 01:14:12 часа, а на момичета е 01:30:39 часа. Подробен анализ на информацията, свързана с продължителността (времетраенето) на записите, е представена в таблица 5.4. На фигура 5.2 е отразено разпределението на времетраенето на всички записи, направени от момичета-момчета, съответно мъже-жени, от различните възрастови групи. От тази фигура се вижда, че най-голям дял заемат записите на 6-годишните момичета диктори, а от момчетата най-голям относителен дял имат момчетата диктори на 5 години.

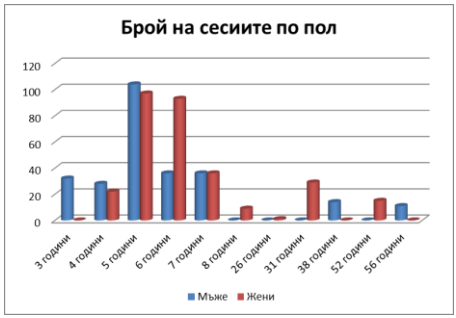
Както бе споменато в предишната глава, една сесия се равнява на един аудиозапис, затова и двете понятия ще бъдат използвани тук като взаимнозаменяеми. Броят на всички сесии, налични в корпуса ChildBG, е 563, от тях 261 са на диктори от мъжки пол, а 302 от женски пол. 493 са записите на говорима детска реч, а 70 – на реч от възрастни диктори. Най-голям е броят на сесиите, проведени с деца на 5 години, следвани от тези с 7-

годишните и т.н. Това съотношение е реално отражение на възрастовото разпределение на децата в предучилищната и 3та група от двете детски градини, които участваха в провеждането на експериментите. 3- и 4-годишните деца са допълнително интервюирани от други детски градини.

Данните за броя на сесиите, разпределени по възраст и пол, са представени в таблица 5.5 и на фигура 5.3.



Фигура 5.2*: Диаграма на разпределението на продължителността на записите по пол на дикторите.*



Фигура 5.3*: Диаграма на разпределението на броя на сесиите по пол на дикторите в ChildBG.*

Всички интервюта са проведени с желанието и съдействието на децата и с разрешението на родителите и на педагогическия състав в учебното заведение.

Като заключение от анализа на данните събрани в корпуса ChildBG може да се каже, че е въведен един иновативен подход при събирането на данни от говорима детска реч, които не са просто записи, а цялостно представяне на емоционалността на децата, семейната среда, тяхното образование и психо-физичните им способности.

5.2 Изследване на измененията на акустични характеристики на говорима детска реч

Изследването на акустичните особености ще бъде обособено в няколко подточки, които изглеждат по следния начин:

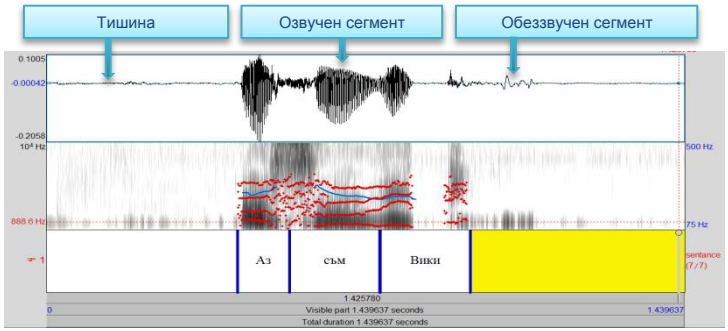
- *Изследване на измененията на фонетичните характеристики на речта за диктор (вътредикторови изменения):* Състои се в изследване на продуцирането на реч от един и същи диктор. Анализират се характеристиките на звуковете на речта, като форманти, продължителност, честоти, височина на основния тон, и тяхното изменение в различните позиции в думите, спрямо ударението и при различни съседни звукове. Изследването е направено за диктори, от различни възрастови групи.
- *Изследване на измененията на фонетичните характеристики между различните диктори в една и съща възрастова група (междудикторови различия):* Състои се в изследване на продуцирането на реч от различни диктори, изследване на няколко фонеми, тяхното изменение в различните позиции в думата и при различни съседни фонеми.

Представените данни са получени след внимателно прослушване на съответните записи, с помощта на програмата Praat, след което те са анализирани и обобщени.

При изследване на речевата вълна се наблюдават три основни елемента [107]:

- **Тишина** (Silance) – Където няма продуциране на реч и допълнителни смущения от артикулационните органи;
- **Обеззвучени** (Unvoiced) – Гласните струни не вибрират, но се забелязват смущения върху спектрограмата, при което се получава звукова вълна с апериодична или произволна природа;
- **Озвучени** (Voiced) – Гласните струни са възбудени и вибрират периодично, така че да се получат полупериодични звукови вълни, под формата на реч.

На фигура 5.4 е представен анализ на спектрограма с ясно обособени периодите на тишина, озвучаване на сигнала, в който се намира същинската информация, и обеззвучения сигнал, който трябва да бъде пропуснат при анализиране. На тази фигура се забелязват основните три форманти F1, F2 и F3 (представени като червени точки), които ще бъдат основен предмет на изследванията в следващите няколко раздела. Освен тях се виждат по-размито четвъртата и петата форманта, но тъй като те не са от значение за вида на продуцираните фонеми, не се разглеждат.



Фигура 5.4: *Спектограма на спонтанна реч на дете на 5 години.*

Периодът на основния тон (pitch) е представен с крива в син цвят и отразява връзката между високите или ниските тонове и броя на трептенията на гласните струни в секунда. Според [144] периода на основния тон е параметърът, носещ информация както за индивидуалността на диктора, така и за емоционалното му състояние. Ето защо той също е обект на настоящите анализи.

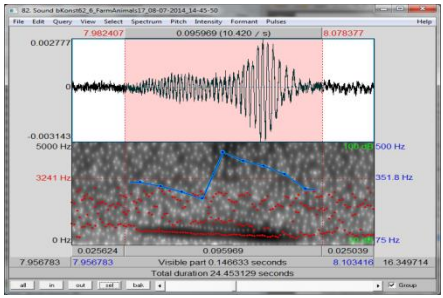
Следващият фактор, който ще бъде изследван, е интензитетът на речевата вълна. Той определя енергията, пренасяна от издишания въздушен поток, и има отношение към оформяне на цялостното представяне на акустичните характеристики на говоримата реч.

При разпознаване на реч друг от основните проблеми е определяне на продължителността на фонемите. Ето защо изследването на продължителността на отделните фонеми е крайно наложително, особено при продуцирането на детска реч.

За изследване на реалната продуцирана реч е необходимо внимателно прослушване и обработване на всеки един от записите. Това е първичното обучение на системите за разпознаване на реч. Освен това е необходимо подготвяне на данните за изграждане на система, основана на знания, за пълноценно и качествено разпознаване на детска реч.

Постоянният (фоновият) шум, който присъства в записите, е с интензитет от 15-17 dB, в зависимост от обстановката, в която е проведено интервюто. Освен този шум, понякога се наблюдават допълнителни шумове с доста по-голям интензитет.

Пример за това е кихането на диктор на 6 години, представено на фигура 5.5. На нея ясно се отличава първата форманта и периода на основния тон. Усреднения интензитет на този звук е 31,5366 dB, което по своята същност не е шум, но няма смислово значение за речта.



Фигура 5.5: *Представяне на спектрограмата на неречевия звук „кихане“ на момче на 6 години.*

Примери за други неречеви звуци се наблюдават при момче bNSK68 на 5 години, по време на изговаряне на думите в колекцията „Отдих и природа“. От здравословна гледна точка, към момента на записване, детето има нарушения в горните дихателни пътища, които участват във фонообразуването. Поради тази причина при продуцирането на някои от думите, то често подсмърча или кашля. Неговото подсмърчане е с усреднен интензитет от 22,7056 dB, а при кашлянето е установен интензитет от 32,3519 dB.

Към настоящия момент на писане на тази дисертация не бяха открити изследвания за неречеви звуци, продуцирани от деца между 4-6 годишна възраст. Например в източник [127] е направено изследване на смях на възрастни диктори и са изведени фонетичнити му характеристики.

От изложеното до тук може да се каже, че изследването на речевия сигнал в реални условия представлява съвкупност от неречеви звуци, постоянен фонов шум и говорима реч.

5.2.1 Изследване на измененията на фонетичните характеристики на реч от един и същи диктор (дете)

В настоящия раздел ще бъдат представени фонетичните изменения, които настъпват при продуцирането на реч от един диктор, приет като представител на изследваните възрастови групи, а именно децата от 4 до 6 години. Подобни анализи са правени за различни езици и те са обобщени в Глава 2.

Тук изследванията са насочени изцяло върху говорима реч на деца, чийто основен (майчин) език е български. Използваните данни са взети от създадения от автора корпус от говорима реч ChildBG, от който са достъпни и записи на диктори, от други възрастови групи.

В потока на свързаната реч е трудно да се установят отделните характеристики на звуковете. От предложения фонетичния модел бе установено, че ударението в българския език е централизиращо и е трудно за моделиране, тъй като неговата позиция е непостоянна. Освен това то е силово и оказва въздействие върху акустичните характеристики на звуковите единици (фони). Ето защо обектът на изследване и анализ ще бъде звукът [ɛ], съответстващ на буква [е], съобразно позицията ѝ под и извън ударение. Тази фонема е избрана като пример за вариране на звук, тъй като според фонетичните

правила на книжовния български език [148] почти не се изменя. В действителност при проведените наблюдения на детска реч този звук често бива изпуснат (елизия на звуковете), асимилиран или акомодиран от неговия съседен. Като пример може да се посочи думата „слонче“ изказана от диктора bAlex33 на 6 години и думите „тигърче“ и „кученце“ изказана от диктор bVGA73 на 4 години, при които звукът [ɛ] в края на думата рязко се отличава по продължителност от останалите произношения на този звук при същите условия.

За представител (представителна извадка) на децата на 6 години е избрано момче с абревиатура bAlex33, за представител на децата на 5 години е избрано момиче с означение gKatrin60, а за представител на децата на 4 години е избрано момче с означение bVGA73. Тъй като получените резултати от останалите диктори от съответните възрасти не се отличават съществено, тук ще бъдат разгледани само тези три деца. Тяхната реч е използвана за определяне на фонетичните изменения, които настъпват при речеобразуването на едни и същи звуци (вътредикторовите различия, от английски inter-speaker temporal variability) за деца на съответната възраст.

В следващите три таблици са представени данните за фонемата [ɛ] и нейното изменение в трите основни местоположения за една гласна: под ударение, извън ударението и в краесловието (в последната сричка). Думите, които са използвани за извличане на звука [ɛ], са от колекцията от думи „Бебета на животни“ от корпуса ChildBG, която се състои от думите: агне, козле, конче, слонче, тигърче, кученце, прасенце, магаренце, мече. Изследваните характеристики са усреднените стойности на първите три форманти, на периода на основния тон, на интензитета и реалната продължителност (времетраене) на продуцирането на звука. За улеснение отделните позиции на звука спрямо ударението са маркирани с различни цветове.

Всички получени характеристики от отделните диктори са представени в таблици 5.6, 5.7 и 5.8. Използваните данни са представени с точност до четвъртия знак след десетичната запетая.

От таблица 5.6 ясно се вижда вариацията на звука [ɛ] при избрания 6-годишен диктор. Обобщеният диапазон на периода на основния тон в краесловието е 301,3474 Hz, извън ударението е 263,0966 Hz и под ударение е 277,0333 Hz. От това може да се заключи, че най-много емоция децата на 6 години влагат в края на думата, макар че се очаква това да бъде на звука, попаднал под ударение.

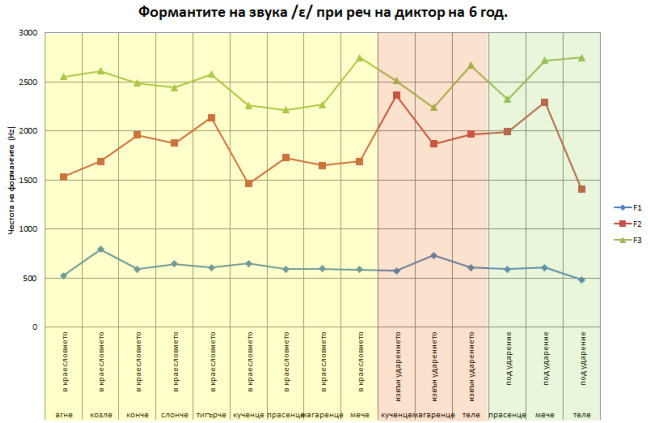
***Таблица 5.6**: Фонетичните различия на звука [ɛ] при диктор bAlex33, на 6 г.*

Звук	Дума	Местоположение в думата	F1	F2	F3	Период на основния тон	Интензитет	Време-траене
ɛ	агне	в краесловието	524,5542 Hz	1532,2850 Hz	2550,5160 Hz	217,6705 Hz	42,2543 dB	0,0855 s
ɛ	козле	в краесловието	790,7443 Hz	1688,3140 Hz	2612,0460 Hz	268,0641 Hz	43,2908 dB	0,1545 s
ɛ	конче	в краесловието	589,1333 Hz	1957,4860 Hz	2487,7390 Hz	304,7139 Hz	45,8019 dB	0,1191 s
ɛ	слонче	в краесловието	645,3247 Hz	1874,8740 Hz	2442,0950 Hz	345,4488 Hz	40,3870 dB	0,0930 s
ɛ	тигърче	в краесловието	607,5827 Hz	2135,1410 Hz	2575,4190 Hz	375,1980 Hz	43,4118 dB	0,1066 s
ɛ	кученце	в краесловието	646,5784 Hz	1460,4880 Hz	2259,7530 Hz	313,6537 Hz	40,8927 dB	0,1555 s
ɛ	прасенце	в краесловието	591,6873 Hz	1726,6000 Hz	2212,5890 Hz	313,0270 Hz	41,1032 dB	0,1343 s
ɛ	магаренце	в краесловието	593,2516 Hz	1649,0399 Hz	2268,5593 Hz	337,4531 Hz	41,5307 dB	0,1013 s
ɛ	мече	в краесловието	584,9198 Hz	1690,2341 Hz	2744,5951 Hz	236,8972 Hz	42,7987 dB	0,1020 s
ɛ	кученце	извън ударение	573,1403 Hz	2363,6149 Hz	2511,8447 Hz	290,7150 Hz	43,1617 dB	0,0833 s
ɛ	магаренце	извън ударение	729,2811 Hz	1865,8010 Hz	2240,2590 Hz	228,6444 Hz	30,0233 dB	0,0350 s
ɛ	теле	извън ударение	605.6380 Hz	1963.8283 Hz	2668.5523 Hz	269.9303 Hz	42.9207 dB	0.0940 s
ɛ	прасенце	под ударение	589,2761 Hz	1989,7460 Hz	2319,9470 Hz	235,6927 Hz	32,6615 dB	0,0976 s
ɛ	мече	под ударение	606,7871 Hz	2293,7579 Hz	2719,2339 Hz	312,4224 Hz	48,3558 dB	0,0988 s
ɛ	теле	под ударение	480.4537 Hz	1407.4989 Hz	2745.0225 Hz	282.9849 Hz	43.9427 dB	0.1952 s

Обобщените стойности на интензитета са както следва: в краесловието е 42,3857 dB, извън ударението е 38.7019 dB и под ударение е 41,6533 dB. От тези резултати може да се направи извод, че децата на 6 години използват повече енергия за завършване на речеобразуването, т.е. в края на думите.

Усреднената продължителност (времетраене) при изговор на звука [ɛ] има следния вид в краесловието е 0,1168 s, извън ударение е 0,0708 s, а под ударение е 0,1305 s. От тези стойности може да се заключи, че при 6-годишните деца най-голямата продължителност на изговор на гласен звук се наблюдава, когато той е под ударение, което не се отличава и от стандартната тенденция в българския език.

Тъй като формантите са най-важните елементи при разпознаването на реч, тяхното разпределение за по-голяма прегледност е представено на фигура 5.6. От нея ясно проличават големите вариации на един и същи звук при едни и същи местоположения в думата, които се проявяват при малките деца. Пример за това са думите кученце, прасенце, магаренце, при които звукът [ɛ] се намира в краесловието и негов съсед е дифтонга [ʧj], а той е предшестван от звука [n]. Тук колебанията са малки.



***Фигура 5.6**: Изследване на изменението на честотите на първите три форманти (F1, F2, F3) при изговор на звука [ɛ] от диктор bAlex33, на 6 г.*

В думите кученце и магаренце звукът [ɛ] извън ударение търпи големи изменения при втората (F2) и третата форманта (F3). От думите се вижда, че десните съседи са еднакви, разликата и влиянието, което се оказва върху изследвания звук, произтича от съседния ляв звук. Както беше представен във фонетичния модел и отразено в акустичното дърво, звукът [ʧ] е шушкав преграден беззвучен звук със средна тоналност, а звукът [rj] е сонорен вибрантен звучен звук с висока тоналност. Съпоставката на двата звука е на ниво 1 от йерархично-фонетичното дърво, което означава, че те са труднозаменяеми в детската реч. Под тяхно влияние честотите на втората и третата форманта на звука [ɛ] претърпяват значителни изменения, въпреки еднаквите си десни съседи.

Като заключение за изменението на формантите при 6-год. диктори може да се каже, че диапазонът на първата форманта варира от 480 Hz до 791 Hz, на втората форманта – от 1407 Hz до 2364 Hz, а диапазона на третата форманта – от 2213 Hz до 2745 Hz. Най-голямо изменение се наблюдава при F2 и то обхваща близо един диапазон (1000 Hz).

Както вече бе споменато за изследване на измененията на речта при един и същи диктор на 5 години, бе избрана реч на момиче с означение gKatrin60 в корпуса от детска реч ChildBG. Всички получени резултати при изследване на вариациите на звука [ɛ], са обобщени в таблица 5.7.

***Таблица 5.7**: Фонетичните различия на звука [ɛ] при диктор gKatrin60, на 5 г.*

Звук	Дума	Местоположение в думата	F1	F2	F3	Период на основния тон	Интензитет	Време-траене
ɛ	агне	в краесловието	475,9501 Hz	1209,9170 Hz	2586,6379 Hz	374,2844 Hz	47,6408 dB	0,2662 s
ɛ	козле	в краесловието	572,1170 Hz	1089,6760 Hz	2688,2995 Hz	287,9220 Hz	44,2163 dB	0,3077 s
ɛ	конче	в краесловието	691,5518 Hz	1389,6066 Hz	2643,7453 Hz	414,3535 Hz	50,5492 dB	0,2093 s
ɛ	слонче	в краесловието	696,2399 Hz	1543,1819 Hz	2721,4621 Hz	182,5339 Hz	43,2689 dB	0,1908 s
ɛ	тигърче	в краесловието	502,2176 Hz	1390,6239 Hz	2703,0700 Hz	168,9439 Hz	40,0351 dB	0,1337 s
ɛ	кученце	в краесловието	553,8185 Hz	1031,4726 Hz	2626,8641 Hz	439,7287 Hz	53,5496 dB	0,1504 s
ɛ	прасенце	в краесловието	710,6652 Hz	1430,9238 Hz	2697,9446 Hz	391,0452 Hz	47,5970 dB	0,2292 s
ɛ	магаренце	в краесловието	547,2974 Hz	1006,2858 Hz	2664,1797 Hz	398,8415 Hz	44,3174 dB	0,1568 s
ɛ	мече	в краесловието	580,7183 Hz	1163,3119 Hz	2701,9570 Hz	396,0975 Hz	45,2287 dB	0,1853 s
ɛ	кученце	извън ударение	412,8319 Hz	965,3414 Hz	2679,0447 Hz	283,8377 Hz	55,6401 dB	0,1108 s
ɛ	магаренце	извън ударение	561,1715 Hz	1629,4252 Hz	2599,5713 Hz	239,5840 Hz	40,1348 dB	0,1566 s
ɛ	теле	извън ударение	504,2416 Hz	888,1314 Hz	2620,6500 Hz	382,3877 Hz	55,6123 dB	0,1475 s
ɛ	прасенце	под ударение	487,5738 Hz	1163,2091 Hz	2735,3517 Hz	213,9154 Hz	43,4087 dB	0,1979 s
ɛ	мече	под ударение	459,3368 Hz	975,2406 Hz	2686,4372 Hz	255,9881 Hz	43,9092 dB	0,2102 s
ɛ	теле	под ударение	571,4766 Hz	1190,7936 Hz	2718,0247 Hz	240,5218 Hz	49,3806 dB	0,1953 s

Обобщеният диапазон на периода на основния тон в краесловието е 339,3056 Hz, извън ударението е 301,9365 Hz и под ударение е 236,8084 Hz. Както при 6-годишния диктор, така и при 5-годишния се наблюдава по-емоционално поведение при продуциране на звука [ɛ], когато се намира в края на думата, и най-малко, когато е под-ударение. Освен това тук се наблюдава понижаване на честотата на основния тон като цяло. Това по-ясно се вижда на фигура 5.8.

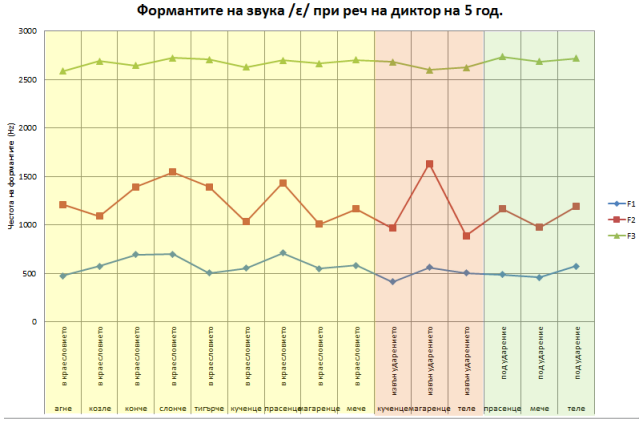
Обобщените стойности на интензитета са както следва: в краесловието е 46,2670 dB, извън ударението е 50,4624 dB и под ударение е 45,5662 dB. За разлика от речта, продуцирана при 6-годишния диктор, тук имаме почти еднаква интонация, което е и причината за близките стойности на интензитета. Сравнението между получените резултати може да се види на фигура 5.9.

Средната продължителност за изговор на звука [ɛ] (фиг. 5.10) при 5-годишното дете в краесловието е 0,2033 s, извън ударението е 0,1383 s, а под ударение е 0,2012 s. От тези стойности ясно проличава закономерността, която е наблюдавана в други подобни изследвания, а именно, че при по-малките деца се наблюдава удължено продуциране на звуковете. Допълнителен получен резултат е, че по-малките деца удължават продуцирането на крайната гласна.

Графичното представяне на формантите на речта на 5-годишния диктор може да се види на фигура 5.7. Тук се наблюдават големи вариации само във втората форманта. Най-ярко отличимо е изменението на звука [ɛ] в краесловието, при думите конче, слонче и тигърче, и в звука извън ударение при думите кученце и магаренце. Втората тенденция бе наблюдавана и при 6-годишния диктор, като за разлика от 5-годишния тук изменението е в трите форманти.

Като заключение за изменението на формантите при 5-годишните диктори може да се каже, че диапазонът на първата форманта варира от 413 Hz до 711 Hz, т.е. с около 50 Hz по-ниско, отколкото при 6-годишния диктор, на втората форманта варира от 888 Hz до 1629 Hz (около 500 Hz по-ниско от това при 6-годишния), а диапазона на третата форманта

варира от 2587 Hz до 2735 Hz, т.е. при F3 се наблюдава намаляване само при минималната стойност, а горната граница почти се запазва. Както и при 6-годишния диктор, най-голямото изменение на честотите на формантите при 5-годишните се наблюдава при F2, като разликата тук е около 740 Hz.



Фигура 5.7: *Изследване на изменението на честотите на първите три форманти (F1, F2, F3) при изговор на звука [ɛ] от диктор gKatrin60, на 5* г.

Внимателното анализиране на аудиозаписа от продуцирана реч върху колекцията от думи „Бебета на животни“ на избрания 4-годишен диктор bVGA73 доведе до получаване на акустичните характеристики на звука [ɛ], представени в таблица 5.8.

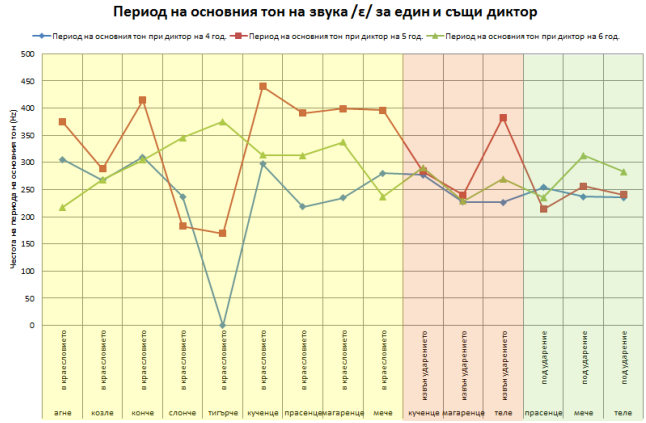
Таблица 5.8: *Фонетичните различия на звука [ɛ] при диктор bVGA73, на 4 г.*

Звук	Дума	Местоположение в думата	F1	F2	F3	Период на основния тон	Интензитет	Продължителност
ɛ	агне	в краесловието	570,0543 Hz	1821,4846 Hz	2466,0461 Hz	305,4961 Hz	38,8957 dB	0,1984 s
ɛ	козле	в краесловието	579,5790 Hz	2250,0979 Hz	2641,4613 Hz	267,1684 Hz	37,8481 dB	0,1475 s
ɛ	конче	в краесловието	546,8707 Hz	1845,6941 Hz	2491,6378 Hz	309,6221 Hz	42,0224 dB	0,1309 s
ɛ	слонче	в краесловието	454,7904 Hz	1814,8541 Hz	2421,2829 Hz	236,7361 Hz	33,0249 dB	0,0730 s
ɛ	тигърче	в краесловието	967,4276 Hz	1986,9901 Hz	2488,7089 Hz	-	21,2109 dB	0,0410 s
ɛ	кученце	в краесловието	494,9531 Hz	1640,7411 Hz	2394,3891 Hz	297,2018 Hz	37,8815 dB	0,0940 s
ɛ	прасенце	в краесловието	539,5606 Hz	1929,4983 Hz	2388,2965 Hz	218,6422 Hz	36,1181 dB	0,1021 s
ɛ	магаренце	в краесловието	544,8869 Hz	1640,9349 Hz	2423,4024 Hz	234,6670 Hz	29,1780 dB	0,1023 s
ɛ	мече	в краесловието	543,6413 Hz	1990,8643 Hz	2435,3680 Hz	280,2627 Hz	35,9162 dB	0,1783 s
ɛ	кученце	извън ударение	531,4626 Hz	2258,5013 Hz	2463,2735 Hz	276,7231 Hz	42,0309 dB	0,0659 s
ɛ	магаренце	извън ударение	384,5007 Hz	1751,0160 Hz	2419,4916 Hz	227,1728 Hz	38,0153 dB	0,0440 s
ɛ	теле	извън ударение	623.3480 Hz	1616.3772 Hz	2671.7069 Hz	226.5485 Hz	30.7486 dB	0.0705 s
ɛ	прасенце	под ударение	526,8297 Hz	1955,4935 Hz	2332,1262 Hz	253,8802 Hz	43,7434 dB	0,0839 s
ɛ	мече	под ударение	580,7049 Hz	1877,6496 Hz	2586,9187 Hz	236,9194 Hz	31,5057 dB	0,0822 s
ɛ	теле	под ударение	653.1502 Hz	1456.6369 Hz	2704.3152 Hz	235.1791 Hz	33.0434 dB	0.1658 s

Усреднените стойности за периода на основния тон (фиг. 5.8) в краесловието е 268,7246 Hz, извън ударението е 243,4815 Hz и под ударение е 241,9929 Hz. Тенденцията, която се наблюдава при 6- и 5-годишните деца, се запазва и тук, че децата са по-емоционални при продуциране на звука [ɛ] в краесловието. Освен това тук се наблюдава понижаване на честотата на основния тон, в сравнение с 5 годишните деца.

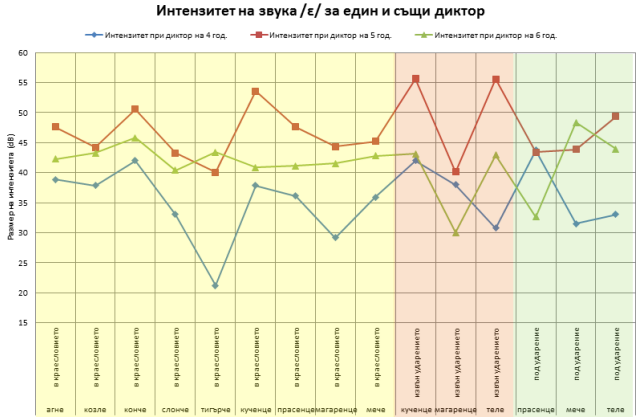
При извеждане на характеристиките на периода на основния тон на гласната [ɛ] при използваната реч на 4-годишния диктор се установи, че в думата тигърче, той не може да бъде изчислен със стандартните алгоритми за векторно квантуване и линейно

предсказване предложени в [144], както и метода на линейна регресия, заложен в Praat. Ето защо тази стойност е неизвестна към настоящия момент на писане на дисертацията.



Фигура 5.8: *Изследване на изменението на периода на основния тон при изговор на звука [ɛ] от диктори на 4, 5 и 6 години*

Това явление е наблюдавано при около 10 % от изследваните диктори. Най-често се среща при звуковете [s, ʦj, ʃ, t] и е характерно при речта на деца от 3-до 7-годишна възраст.

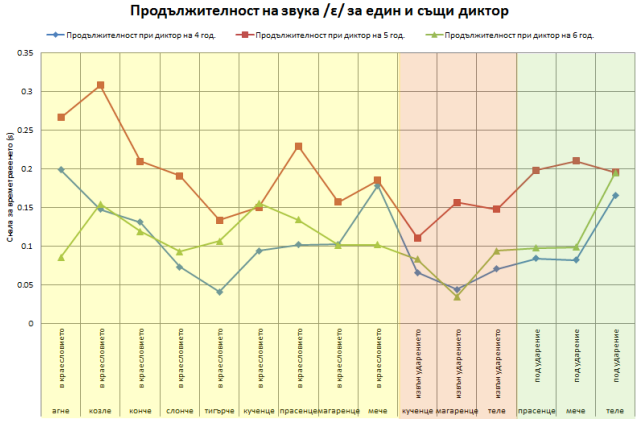


Фигура 5.9: *Изследване на осреднената стойност на енергийния интензитет (силата на гласа) при изговор на звука [ɛ] от диктори на 4, 5 и 6 години*

Обобщените стойности на интензитета (фиг. 5.9) са в краесловието 34,6773 dB, извън ударението 36,9316 dB и под ударение 36,0975 dB. Интересното, което тук се получава като резултат е, че интензитетът се запазва независимо от местоположението на изговорения звук [ɛ]. Това ясно контрастира с наблюдаваните вариации в интензитета при 5- и 6-годишните деца.

Средната продължителност за изговор на звука [ɛ] при 4-годишното дете (фиг. 5.10) в краесловието е 0,1186 s, извън ударението е 0,0601 s, а под ударение е 0,1106 s.

Резултатът може да се обясни като прецедент, при който 4-годишно дете говори по-бързо от 5-годишно. Това се обяснява с малката разлика между двете деца, семейната среда и психологичните особености, които са важни фактори за развитието на говорно-комуникативните способности на децата в тази възраст. Това е пример за нарушение на правилото, че при по-малките деца се наблюдава по-голямо удължаване на речевите звуци. От графиката представена на фигура 5.10, се вижда, че използваната реч на 4-годишния диктор се доближава като времетраене до тази на речта на 6-годишния. Това е повлияно и от експерта, тъй като изследваните записи на 4-и 6-годишните са записани с първия описан метод, а именно с помощта на ръководител и децата в този случай са имитирали дикцията му. Оттук произтича и този дисбаланс в получените резултати.



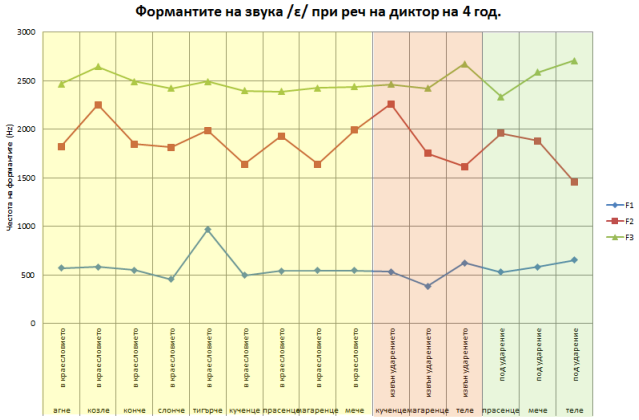
***Фигура 5.10**: Изследване на изменението на времетраенето (продължителността) на изговор на звука [ɛ] от диктори на 4, 5 и 6 години.*

Графичното представяне на формантите на речта на 4-годишния диктор може да се види на фигура 5.11. За разлика от формантите на предходните двама диктори, тук се наблюдават вариации в стойностите на всички. Както при 5-годишния диктор, така и тук има големи изменения на звука [ɛ] в краесловието, при думите слонче и тигърче, и в звука извън ударение при думите кученце и магаренце. Наблюдаваната тенденцията при 5-годишния диктор, за изменение в трите форманти при тези думи, се запазва и за 4-годишния.

Като заключение за изменението на формантите при 4 годишните диктори може да се каже, че диапазонът на първата форманта варира от 384 Hz до 967 Hz. При получените изменения бе установено, че най-ниската стойност е с около 30 Hz под най-ниската за 5-годишния, а най-високата – с около 200 Hz по-висока отколкото при 5-годишния диктор. Диапазонът при втората форманта варира от 1457 Hz до 2258 Hz (почти аналогични стойности като тези на 6-годишния диктор), а диапазонът на третата форманта - от 2332 Hz до 2704 Hz. Отчетени са изменения на първата, втората и третата форманта.

От направените изследвания, свързани с продължителността на продуциране на звуковите единици в българския език, бе установено, че с най-малка продължителност са звуците [t, n, m], а с най-голяма продължителност са [ɔ, a]. Най-лесно различими от спектрограмата са [ʦ, ʧ, ʃ], а най-трудно [l] и потъмнените гласни [o, e].

Освен това, когато продуцираната реч от деца между 4 и 6 години е повлияна от дикцията, интонацията и тона на експерта, който е произнесъл думата, тогава се наблюдават постоянство и устойчивост на речевите им характеристики.



Фигура 5.11: *Изследване на изменението на честотите на първите три форманти (F1, F2, F3) при изговор на звука [ɛ] от диктор bVGA73, 4 г.*

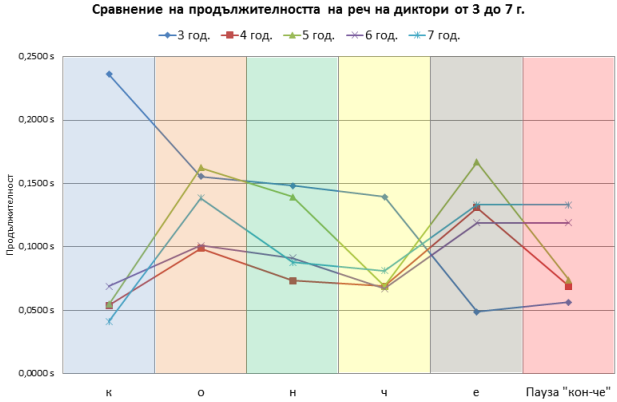
С проведения експеримент бе доказано, че имитирането на начина на продуциране на реч на експерта по време на записване на говорима детска реч води до изменения на акустичните характеристики и невъзможността за реалното отразяване на речевите способности на децата от 4 до 6 години, като тази тенденция се запазва и за 3- и 7-годишните.

5.2.2 Изследване на измененията на фонетичните характеристики на реч при различни диктори (деца)

Голям интерес представлява и определянето на границите на отделните срички в думата. С други думи трябва да се изследва времето за поемане на дъх при изговора на отделните срички в една дума. Често алгоритмите за преемфазис използват интензитета на постоянния фонов шум, и чрез неговото премахване от основния сигнал, получават чистия речеви сигнал. Проблемът, който се появи при използването на стандартно филтриране с преемфазис беше, че поради високите колебания в интензитета на гласа на децата бяха редуцирани и сигнали със смислово значими стойности (озвученият сигнал). Това се получава тъй като често говоримата реч попада под прага или е на самия граничен праг на изследвания интервал. Пример за това е изказването на диктор bGogo56 (3 години) на думата „МАГАРЕНЦЕ“, при която се наблюдава фонов шум с усреднен интензитет от 15 dB, а последният звук [ɛ] се произнася с усреднен интензитет (силата на гласа) от 17 dB. След тестване със стандартното филтриране чрез преемфазис с помощта на Praat, гласната [ɛ] е изрязана и остава само омекотеното „ц“, което фонетично се означава с [tsʲ]. Така получената думата е „МАГАРЕНЦ“ и губи първоначалния си смисъл.

За отразяване на различията между отделните диктори при продуцирането на едни и същи звуци бяха изследвани 30 деца на възраст между 3 и 7 години. Разпределението на дикторите е както следва: 2 деца на 3 год., 3 деца на 4 год., 13 деца на 5 год., 7 деца на 6 год. и 4 деца на 7 год. Звуците, които бяха изследвани са к - [k], о - [ɔ], н - [n], ч - [ʧ] и е - [ɛ],

при изговор на думата „конче“. На фигура 5.12 и в таблица 5.9 са отразени усреднените продължителности при тяхното речеобразуването.



Фигура 5.12 *Графично представяне на продължителността на звуковете в думата „конче“ на диктори от пет различни възрастови групи (от 3 до 7 год.).*

На графиката на фигура 5.12 се наблюдават най-големи изменения при 3-годишните диктори, като за съгласните [k] и [ʧ] има най-голямо препокриване в продължителността на продуциране.

Таблица 5.4: *Продължителността на отделните звукове в думата „конче“ на диктори от пет различни възрастови групи (от 3 до 7 год.).*

		Възраст				
		3 год.	**4 год.**	**5 год.**	**6 год.**	**7 год.**
Продължителност	**к**	0,2360 s	0,0538 s	0,0549 s	0,0691 s	0,0413 s
	о	0,1555 s	0,0989 s	0,1626 s	0,1011 s	0,1386 s
	н	0,1482 s	0,0733 s	0,1395 s	0,0915 s	0,0879 s
	ч	0,1395 s	0,0690 s	0,0691 s	0,0673 s	0,0814 s
	е	0,0490 s	0,1309 s	0,1670 s	0,1191 s	0,1333 s
	Пауза "кон-че"	0,0563 s	0,0690 s	0,0739 s	0,1191 s	0,1333 s

При 4- и 5-годишните деца има почти пълно съвпадение в усреднените стойности на паузата между отделните срички, като стойността при 3 годишните също клони към този диапазон от около 0,0664 s. Това което прави впечатление е, че при по-големите деца времето за поемане на дъх между отделните срички е по-голямо и надвишава 10 ms, докато при по-малките деца сричкоотделянето е по-труден процес и трудно се отличава.

5.3 Класификация на акустичните характеристики на детска реч

С помощта на модификация на класическия алгоритъм ИСОМАД, който бе представен в предходните глави, ще се изследват двойките форманти (F1, F2) и (F2, F3) с цел класифициране на усреднените стойности на първите три форманти за съответните звукове при различните диктори. С ИСОМАД бързо се определят класовете (клъстерите), към които принадлежат продуцираните звукове. Освен това този алгоритъм позволява анализиране на промените на акустично ниво, предизвикани от специфичното речеобразуване при децата. Това означава, че лесно може да се открият асимилациите,

елазиите, епентезията и т.н., които са характерни за детската реч, и да позволи тяхното тълкуване. Получените резултати са извлечени от програмния продукт разгледан в Глава 4.

Представени са изследвания свързани с продуцирането на думата „телѐ“ от 7 различни диктора. Двама от тях са на 3 години, един на 4 години, двама на 5 години, и по един на 6 и 7 години.

Зададените стойности на параметрите на алгоритъма са:

- Параметъра за очаквания брой на клъстерите: $K = 4$;
- Параметъра определящ максималния брой итерации: $I = 10$;
- Параметъра определящ максималното стандартно отклонение за всеки клъстер: $\theta_s = 0.05$;
- Параметъра определящ минималното разстояние между два центъра на клъстери: $\theta_c = 50Hz$.
- Коефициента за точността: k= $0{,}025$;
- Центрове на клъстери: Определят се от усреднените стойности на формантите на изследваните звукове.

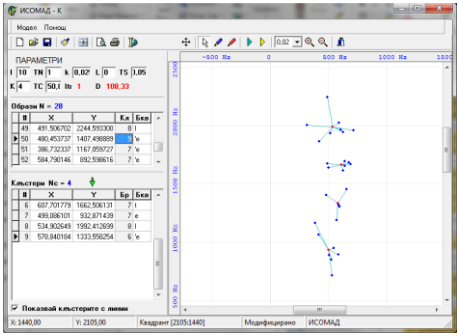
При първата стъпка от класифицирането се разглежда двойката (F1, F2), а на следващата стъпка - двойката (F2, F3) (табл. 5.10).

Таблица 5.5: *Класификация на звуковете от думата „теле“ с помощта на модифицирания алгоритъм ИСОМАД*

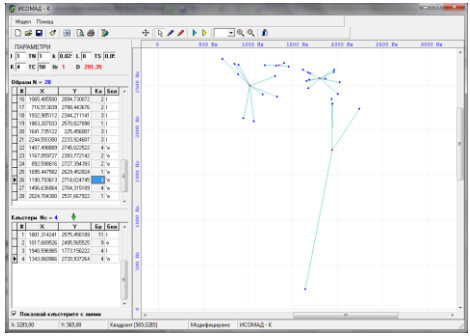
Диктор	Год.	Очакван звук	F1	F2	Получен звук за (F1, F2)	F2	F3	Получен звук за (F2, F3)
bAlex33	6	t	487,7257 Hz	1317,3145 Hz	ɛ	1317,3145 Hz	2716,3866 Hz	'ɛ
bBojidar39	5		585,4764 Hz	1322,9553 Hz	ɛ	1322,9553 Hz	2372,1635 Hz	ɛ
bDBK27	3		690,6587 Hz	1669,3625 Hz	t	1669,3625 Hz	2641,1502 Hz	t
bGogo56	3		640,1888 Hz	1653,8303 Hz	t	1653,8303 Hz	2546,7346 Hz	t
gKatrin60	5		713,6726 Hz	1960,5095 Hz	l	1960,5095 Hz	2645,9362 Hz	t
bVGA73	4		674,3293 Hz	1683,5076 Hz	t	1683,5076 Hz	2559,2773 Hz	t
gIvet57	7		588,3858 Hz	1677,2826 Hz	t	1677,2826 Hz	2464,7533 Hz	t
bAlex33	6	ɛ	605,6380 Hz	1963,8283 Hz	l	1963,8283 Hz	2668,5523 Hz	t
bBojidar39	5		543,1347 Hz	950,5494 Hz	ɛ	950,5494 Hz	2128,7255 Hz	ɛ
bDBK27	3		525,5623 Hz	849,7623 Hz	ɛ	849,7623 Hz	2735,3005 Hz	ɛ
bGogo56	3		594,7584 Hz	1306,1503 Hz	ɛ	1306,1503 Hz	2598,9100 Hz	ɛ
gKatrin60	5		504,2416 Hz	888,1314 Hz	ɛ	888,1314 Hz	2620,6500 Hz	ɛ
bVGA73	4		623,3480 Hz	1616,3772 Hz	t	1616,3772 Hz	2671,7069 Hz	t
gIvet57	7		626,5013 Hz	1943,0743 Hz	l	1943,0743 Hz	2289,0331 Hz	l
bAlex33	6	l	316,5613 Hz	2006,3997 Hz	l	2006,3997 Hz	2399,9585 Hz	t
bBojidar39	5		417,0431 Hz	1065,4855 Hz	ɛ	1065,4855 Hz	2094,7301 Hz	ɛ
bDBK27	3		532,0986 Hz	716,5130 Hz	ɛ	716,5130 Hz	2788,4437 Hz	ɛ
bGogo56	3		514,0475 Hz	1932,9851 Hz	l	1932,9851 Hz	2344,2111 Hz	l
gKatrin60	5		419,3856 Hz	1863,2070 Hz	l	1863,2070 Hz	2570,8279 Hz	t
bVGA73	4		497,6712 Hz	1641,7351 Hz	t	1641,7351 Hz	225,4560 Hz	l
gIvet57	7		491,5067 Hz	2244,5933 Hz	l	2244,5933 Hz	2233,9246 Hz	l
bAlex33	6	'ɛ	480,4537 Hz	1407,4989 Hz	'ɛ	1407,4989 Hz	2745,0225 Hz	'ɛ
bBojidar39	5		386,7323 Hz	1167,0597 Hz	'ɛ	1167,0597 Hz	2393,7721 Hz	'ɛ
bDBK27	3		584,7901 Hz	892,5986 Hz	'ɛ	892,5986 Hz	2727,3944 Hz	'ɛ
bGogo56	3		639,3306 Hz	1695,4476 Hz	t	1695,4476 Hz	2629,4528 Hz	t
gKatrin60	5		571,4766 Hz	1190,7936 Hz	'ɛ	1190,7936 Hz	2718,0247 Hz	'ɛ
bVGA73	4		653,1502 Hz	1456,6369 Hz	'ɛ	1456,6369 Hz	2704,3152 Hz	'ɛ
gIvet57	7		591,9082 Hz	2024,7043 Hz	l	2024,7043 Hz	2531,6679 Hz	t

Първоначалните центрове на клъстерите (класовете) са определени като усреднена стойност на всички данни към първата и втората форманта за един звук. След стартиране на алгоритъма бяха разпознати (класифицирани) 17 звука и сгрешени 11 звука. Графиката на клъстерирането на отделните звукове е представена на фигура 5.12.

От направеното изследване лесно се установява коартикулацията, която оказва силно въздействие върху продуцирането на детската реч. При диктора bAlex33 (6 години) звукът [t] е произнесен със задно учленяване, така че неговото звучене отива по-скоро към гласната [ɛ], а гласната [ɛ] от своя страна е силно омекотена от последвалия звук [l] и учленяването й преминава към устната кухина. От лингвистична гледна точка може да се каже, че при този диктор за първите два звука се наблюдава явлението дисимилация. Останалите два звука от изследваното изказване са разпознати.



Фигура 5.13: Графика, получена след клъстерирането на звуковете по първите две форманти.



Фигура 5.134*: Графика получена след клъстерирането на звуковете съгласно втората и третата форманта.*

При диктора bBojidar39 (5 години) звукът [ɛ] е силно централизиращ и променя артикулацията на неговите съседи, в резултат, на което след класифицирането, те са разпознати именно като него. Това е позиционна промяна, определена от силното поставяне на ударението върху звука [ɛ] и последвалата елизия и асимилация на останалите гласни. За диктора bDBK27 (3 години) се наблюдава силно омекотено звучене на звука [l], при което той е класифициран като [ɛ].

В резултат от класифицирането е получена степен на грешка WER=0,39.

При изследване на втората и третата форманта броят на правилно разпознатите (класифицирани) звуци отново е 17, а броят на сгрешените е 11. Степента на грешка и тук е WER=0,39. Графиката, която е получена в резултат на приложения алгоритъм, е представена на фигура 5.13.

При прилагане на модифицирания алгоритъм ИСОМАД върху данните, получени при изговора на думата „*конче*“, се получава по-малка степен на грешка. При изследване на първата двойка форманти (F1, F2) нейната стойност е WER=0,225, а при двойката (F2, F3) WER=0,15. При изследване на акустичните характеристики на дикторите са използвани същите стойности за параметрите, както при думата „*теле*“.

При изговор на думата „*конче*“ се наблюдават по-малко нарушения, като асимилация и елизиция, в продуцираните звуци. По-големи отклонения има в първите две форманти на изследваните звуци. Звукът [k] е с голяма вариативност при различните диктори и в около 27% от случаите се проявява частична асимилация от гласната [ɔ]. В около 67% от случаите при 3-годишните диктори [ʧ] е силно дисимилирано от съседната гласна [ɛ] и изцяло променя акустичните си характеристики, при което се превръща в [ɛ].

Таблица 5.11: *Класификация на звуковете от думата „конче“ с помощта на модифицирания алгоритъм ИСОМАД*

Диктор	Год.	Очакван звук	F1	F2	Получен звук за (F1, F2)	F2	F3	Получен звук за (F2, F3)
bAlex33	6	k	549,9590 Hz	1248,6580 Hz	ɔ	1248,6580 Hz	2517,2377 Hz	k
bBojidar39	5		719,3683 Hz	1476,2970 Hz	k	1476,2970 Hz	2350,6053 Hz	k
bDBK27	3		554,8343 Hz	1418,3831 Hz	k	1418,3831 Hz	1928,0000 Hz	ɔ
bDBK27	3		520,3521 Hz	1469,6723 Hz	ɔ	1469,6723 Hz	1954,2632 Hz	k
bGogo56	3		830,8025 Hz	1455,3644 Hz	k	1455,3644 Hz	2321,9247 Hz	k
gKatrin60	5		1027,0981 Hz	1604,9492 Hz	k	1604,9492 Hz	2242,4328 Hz	k
bVGA73	4		613,5581 Hz	1235,3066 Hz	ɔ	1235,3066 Hz	2334,5778 Hz	k
gIvet57	7		883,3514 Hz	1439,5401 Hz	k	1439,5401 Hz	2310,4125 Hz	k
bAlex33	6	ɔ	626,7870 Hz	1369,7460 Hz	ɔ	1369,7460 Hz	1672,9973 Hz	ɔ
bBojidar39	5		560,0623 Hz	1291,0552 Hz	ɔ	1291,0552 Hz	1940,9829 Hz	ɔ
bDBK27	3		754,1185 Hz	1279,4992 Hz	ɔ	1279,4992 Hz	1665,0616 Hz	ɔ
bDBK27	3		681,0150 Hz	1269,8105 Hz	ɔ	1269,8105 Hz	1770,6939 Hz	ɔ
bGogo56	3		663,0247 Hz	1420,5188 Hz	ɔ	1420,5188 Hz	1732,9282 Hz	ɔ
gKatrin60	5		710,9523 Hz	1299,1230 Hz	ɔ	1299,1230 Hz	1758,8855 Hz	ɔ
bVGA73	4		523,7349 Hz	1390,2349 Hz	ɔ	1390,2349 Hz	2092,8152 Hz	ɔ
gIvet57	7		667,5467 Hz	1333,4858 Hz	ɔ	1333,4858 Hz	1927,4227 Hz	ɔ
bAlex33	6	n	349,8260 Hz	1525,3591 Hz	n	1525,3591 Hz	2424,7456 Hz	n
bBojidar39	5		348,2368 Hz	1696,7581 Hz	n	1696,7581 Hz	2428,6959 Hz	n
bDBK27	3		434,0888 Hz	1695,8809 Hz	n	1695,8809 Hz	2244,6931 Hz	n
bDBK27	3		408,9419 Hz	1712,2399 Hz	n	1712,2399 Hz	2464,6508 Hz	ɛ
bGogo56	3		386,7246 Hz	1330,6703 Hz	ɔ	1330,6703 Hz	2360,3087 Hz	n
gKatrin60	5		449,1858 Hz	1754,9164 Hz	n	1754,9164 Hz	2221,4884 Hz	n
bVGA73	4		313,2632 Hz	1595,8851 Hz	n	1595,8851 Hz	2482,4782 Hz	n
gIvet57	7		412,6040 Hz	1973,3669 Hz	n	1973,3669 Hz	2332,7623 Hz	ʧ
bAlex33	6	ʧ	736,8260 Hz	2057,9497 Hz	ʧ	2057,9497 Hz	2667,0797 Hz	ʧ
bBojidar39	5		875,1596 Hz	2185,5159 Hz	ʧ	2185,5159 Hz	2556,7347 Hz	ʧ
bDBK27	3		793,8511 Hz	1742,6902 Hz	ɛ	1742,6902 Hz	2588,6486 Hz	ɛ
bDBK27	3		745,3426 Hz	1711,6760 Hz	ɛ	1711,6760 Hz	2659,7648 Hz	ɛ
bGogo56	3		887,7975 Hz	2231,2357 Hz	ʧ	2231,2357 Hz	2634,6470 Hz	ʧ
gKatrin60	5		746,2795 Hz	2156,2654 Hz	ʧ	2156,2654 Hz	2747,3416 Hz	ʧ
bVGA73	4		884,1393 Hz	1796,4315 Hz	ʧ	1796,4315 Hz	2427,1067 Hz	ʧ
gIvet57	7		737,8172 Hz	2100,3080 Hz	ʧ	2100,3080 Hz	2519,5115 Hz	ʧ
bAlex33	6	ɛ	589,1333 Hz	1957,4858 Hz	ʧ	1957,4858 Hz	2487,7394 Hz	ʧ
bBojidar39	5		500,5208 Hz	1390,2952 Hz	ɔ	1390,2952 Hz	2421,4241 Hz	k
bDBK27	3		746,4181 Hz	1703,7607 Hz	ɛ	1703,7607 Hz	2693,2271 Hz	ɛ
bDBK27	3		653,0882 Hz	1729,7254 Hz	ɛ	1729,7254 Hz	2476,1716 Hz	ɛ
bGogo56	3		863,0794 Hz	2151,4625 Hz	ʧ	2151,4625 Hz	2619,9808 Hz	ɛ
gKatrin60	5		691,5518 Hz	1389,6066 Hz	k	1389,6066 Hz	2643,7453 Hz	ɛ
bVGA73	4		546,8707 Hz	1845,6941 Hz	ɛ	1845,6941 Hz	2491,6378 Hz	ɛ
gIvet57	7		537,2109 Hz	1786,9248 Hz	ɛ	1786,9248 Hz	2438,3085 Hz	ɛ

5.4 Изводи

- Представени бяха резултатите от стандартното събиране на говорима детска реч, т.е. с помощта на педагог-логопед, който в игрова форма произнася желаната дума и детето повтаря след него.
- Представени бяха резултатите от използването на предложената мултимедийна система за управление на корпус от говорима детска реч, която предоставя възможност за използване на интерактивни методи в процеса на записване на говорима детска реч.
- Обобщени бяха резултатите от двата начина на събиране на детска реч, качеството на направените записи и тяхната използваемост в процесите за класифициране и разпознаване на реч.
- Анализирани бяха измененията на акустично-фонетичните характеристики в различните възрастови групи на диктори от 4 до 6 години.

- С помощта на модификацията на ИСОМАД бе показан метод за класифициране на акустично-фонетичните особености на говоримата детска реч, и определяне на ефекта от коартикулацията, както на отделните диктори, така и на група от диктори.

ЗАКЛЮЧЕНИЕ

С настоящата дисертация бяха поставени основите на едно задълбочено интердисциплинарно научно изследване с практико-приложен характер:

- Бяха изследвани модерните тенденции в теорията за разпознаване на реч. Направен беше анализ на основните постановки използвани за акустично-фонетично моделиране на детска реч.
- Разгледани бяха специфичните особености на българската фонетика в резултат, на което беше построено йерархично-фонетично дърво, за определяне на уникалната позиция на всяка фонема, съгласно съществуващите правила на българската фонетика.
- Бяха разгледани по-популярни корпуси от говорима детска реч, въз основа на което бе разработен концептуален модел за изграждането на корпус от детска реч на български език.
- Беше разработена интерактивна мултимедийна система за събиране, обработване и анализ на данни съхранявани в корпус от говорима реч на деца от 4- до 6-годишна възраст.
- Изследвани бяха акустичните характеристики на говоримата детска реч. Анализирани бяха специфичните акустично-фонетични изменения, както за един и същи диктор, така и различията между отделните диктори.
- Беше модифициран класическият алгоритъм за клъстеризация ИСОМАД, който беше използван за класификация на звуковете на речта на деца на възраст от 3 до 7 години, и бяха представени получените степени на грешка. С негова помощ бе определен ефекта от коартикулацията, както на отделни диктори, така и за група от диктори.

Насоки за бъдеща работа

- Изследване на акустичните особености на реч на деца с говорно комуникативни нарушения и на деца с диалектно произношение;
- По-задълбочени изследвания на влиянието на външните фактори (семейната среда, училище, допълнителни занимания и т.н.) върху акустично-фонетичните особености на детската реч;
- Изследване на реч на деца извън разглеждания възрастов диапазон в дисертацията;
- Подобряване на предложения модел за автоматично транскрибиране.

ПУБЛИКАЦИИ ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД

1. **Kraleva, R**., Kralev, V. (2006) On modification of the iterative self-organizing data analysis technique algorithm, Journal of the Technical University at Plovdiv "Fundamental Sciences and Applications", Vol. 13 (1) 2006, pp. 121-128

2. **Kraleva, R.** (2009) On model of information system for management of information flows, Union of Scientists in Bulgaria, Section: Blagoevgrad, Yearbook "Science-Education-Art", 2009, Vol. (3), pp. 203-210

3. **Kraleva, R.**, Kralev, V. (2009) On model architecture for a children's speech recognition interactive dialog system, In proc. of the 3th International Scientific Conference "Mathematics and Natural Sciences", Vol. (1), pp. 106-111

4. **Kraleva, R.** (2011) Design and development a children's speech database, In proc. of the 4th International Scientific Conference "Mathematics and Natural Sciences", Bulgaria, Vol. (2), pp. 41-48

5. **Kraleva, R.** (2011) Research modern corpora for automatic children's speech recognition, Union of Scientists in Bulgaria, section: Blagoevgrad, yearbook, 2011, pp. 183-189

6. **Kraleva, R.** (2011) Research and analysis of the difference between children's speech and adults' speech in automatic speech recognition systems, Union of Scientists in Bulgaria, section: Blagoevgrad, yearbook, 2011, pp. 166-175

7. Kralev, V., **Kraleva, R.**, Botseva, D., Kostadinova, D. (2014) On Some Grammatical Aspects of the Speech of Children with Communicative Disorders, Journal "Linguistic World" ("Orbis Linguarum"), South-West University, Vol. 12 (2), pp. 40-42, Bulgaria

8. **Кралева, Р.** (2014) Изследване на начините за събиране на данни в интерактивен мултимедиен корпус от детска реч на български език ChildBG, Сборник от публикации „Електронни форми на обучение в университетското образование", изд. „Авангард Прима", София, стр. 67-78

9. **Кралева, Р.** (2014) Интерактивен мултимедиен корпус от детска говорима реч на български език ChildBG, Сборник от публикации „Електронни форми на обучение в университетското образование", изд. „Авангард Прима", София, стр. 129-143

ПРИНОСИ НА ДИСЕРТАЦИОННИЯ ТРУД

1. Въз основа на резултатите от направеното проучване и сравнителния анализ между различни корпуси от говорима реч е предложен концептуален модел и е представена архитектурна схема на информационна система за събиране и обработване на данни от говорима реч. Формулирани са основните функционални изисквания към системите от този клас.

2. В съответствие с избрания модел на данни е проектирана и създадена релационна база от данни, която да съхранява необходимата информация, свързана със събирането и обработката на данните от говорима реч;

3. Проектирана е и е създадена интерактивна мултимедийна система за събиране, обработване и анализ на данни от говорима реч. Разработената система е изследвана и е доказана нейната работоспособност. Чрез нея е попълнен с данни корпус от говорима детска реч на български език на диктори от 3 до 8 години.

4. Предложена е модификация на "Итеративния СамоОрганизиращ се Метод за Анализ на Данни" – ИСОМАД (Iterative Self Organizing Data Analysis Techniques Algorithm – ISODATA). В сравнение с класическия алгоритъм е постигнато подобрение при разпределянето на образите по клъстери, като експериментално е доказана ефективността на модифицирания алгоритъм.

5. Предложен е контекстно зависим фонетичен модел за транскрибиране на български език, базиран на трифонния модел. За целта е създадено и йерархично-фонетично дърво на български език. Също така е разработена и уеб базирана информационна система, в която е имплементиран предложеният алгоритъм.

6. Експериментално бяха изследвани акустичните характеристики на говоримата детска реч. Анализирани бяха специфичните акустично-фонетични изменения, както за един и същи диктор, така и различията между отделните диктори. Получените резултати от експериментите са анализирани и обобщени. Използван е също така модифицираният алгоритъм ИСОМАД за класифициране на акустичните характеристики на детската реч.

Формулираните приноси имат значение в научен, в научно-приложен и в приложен аспект, както следва: в научен аспект - приноси 4 и 5, в научно-приложен аспект: приноси 1 и 6, и в приложен аспект: приноси 2 и 3.

Приносите са апробирани в следните публикации: принос 1 – [2], [3], [5]; принос 2 - [4]; принос 3 - [9]; принос 4 - [1]; принос 6 – [6], [7] [8].