СИНТЕЗ И РАСПОЗНАВАНИЕ РЕЧЕВОГО СИГНАЛА С ПРИМЕНЕНИЕМ УНИФИЦИРОВАННОГО ЯЗЫКОВОГО ПРЕДСТАВЛЕНИЯ

Калимолдаев М. Н., Мусабаев Р.Р.

Институт проблем информатики и управления МОН РК,
Алматы, Казахстан
e-mail: rmusab@gmail.com

На современном этапе развития информационных технологий стоит задача их интеллектуализации [1]. Решение данной задачи позволит перевести взаимодействие человека и машины на качественно новый уровень. Одним из основных направлений интеллектуализации сферы информационных технологий является интенсификация применения передовых человеко-машинных интерфейсов. Наиболее интуитивными для человека являются речевые интерфейсы взаимодействия [2]. Для их реализации используются технологии синтеза и распознавания речи.

Для качественного решения сложных задач в определённой проблемной области наиболее оптимальным методом является разработка и использование специализированных языков, с помощью которых можно осуществить формализацию задачи, описывать обрабатываемые данные, а также задавать наборы инструкций и команд. В свою очередь для использования специализированных языков необходимо осуществлять их автоматический синтаксический анализ [3] и последующую компиляцию.

Специально для решения задач синтеза и распознавания речевого сигнала [4] разработан унифицированный язык фонетического представления (Unified Phonetic Language - UPL). Фактически данный язык является расширенной фонетической транскрипцией. На языке UPL описываются свойства речевого сигнала, по которым осуществляется его синтез либо распознавание. В случае синтеза речи на языке UPL описываются требуемые характеристики речевого сигнала, на основе которых компилятор выбирает наиболее подходящие элементы компиляции и осуществляет последующую генерацию сигнала. В системах распознавания речи у исходного речевого сигнала измеряются основные характеристики, осуществляется распознавание отдельных фонем с последующей генерацией языкового описания распознанных сегментов на языке UPL. При этом более высокоуровневые подсистемы осуществляют последующее распознавание отдельные слов по их UPL-представлению.

В Институте проблем информатики и управления МОН РК ведутся научно-исследовательские работы по разработке методов и алгоритмов синтеза устной казахской речи [5], которые в настоящее время имеют законченную программную реализацию на уровне синтезатора казахской речи по фонетическому UPL-представлению. На текущем этапе разработки стоит задача доведения качества синтезированного речевого сигнала до максимального уровня.

В качестве основных свойств описываются следующие параметры синтезируемого речевого сигнала:

1. Фонетическая транскрипция с указанием ударений на гласных звуках (0 – безударный; 1 – основное ударение; 2 – второстепенное ударение).
2. Указание длительностей фонем и пауз в миллисекундах.
3. Указание номеров аллофонов - акустических реализаций фонем (одного из различных вариантов произношения).
4. Задание огибающей частоты основного тона в виде перечисления координат опорных точек кривой Безье в плоскости положение-частота.

5. Задание огибающей амплитудного уровня в виде перечисления координат опорных точек кривой Безье в плоскости положение-амплитуда.

Синтезируемый сигнал на унифицированном языке фонетического представления языке описывается в следующем виде:

Фонема1Ударение1(Длительность1;Аллофон1;[ЧОТ1];{Амплитуда1})
Фонема2Ударение2(Длительность2;Аллофон2;[ЧОТ2];{Амплитуда2})
...
ФонемаNУдарениеN(ДлительностьN;АллофонN;[ЧОТN];{АмплитудаN}),

где *ФонемаN* – мнемоническое обозначение фонемы, *УдарениеN* – признак ударения, *ДлительностьN* – значение длительности звучания фонемы в миллисекундах, *АллофонN* – номер аллофонной реализации фонемы, *ЧОТN* и *АмплитудаN* – соответственно контура частоты основного тона и амплитуды, которые задаются следующим образом: $X_1, Y_1; X_2, Y_2; ...; X_m, Y_m$, где X_m – относительная координата *m*-ой опорной точки на отрезке от начала (0) и до конца (1) звучания фонемы $X_m \in [0;1]$, Y_m - значение частоты основного тона ($X_m \in [0;+\infty]$) либо амплитуды ($X_m \in [0;1]$). Приведём пример описания казахского слова «бала» на унифицированном языке фонетического представления:

PAU(160;1)
B(43;1199;[*0,98;0.534,99.46;1,101.28*];{*0,0;0.6,0.1;1,0.2*})
A0(82;1;[*0,101.28;0.5,106;1,103.46*];{*0,0.2;0.5,0.21;1,0.2*})
L(78;1;[*0,102.92;0.452,104.38;1,106.74*];{*0,0.2;0.5,0.1;1,0.2*})
A1(127;1184;[*0,106.74;0.473,108.81;1,102.03*];{*0,0.2;0.5,0.21;1,0*})
PAU(160;1).

На Рис. 1 приведен результат компиляции данного UPL-представления в системе синтеза речевого сигнала на казахском языке.

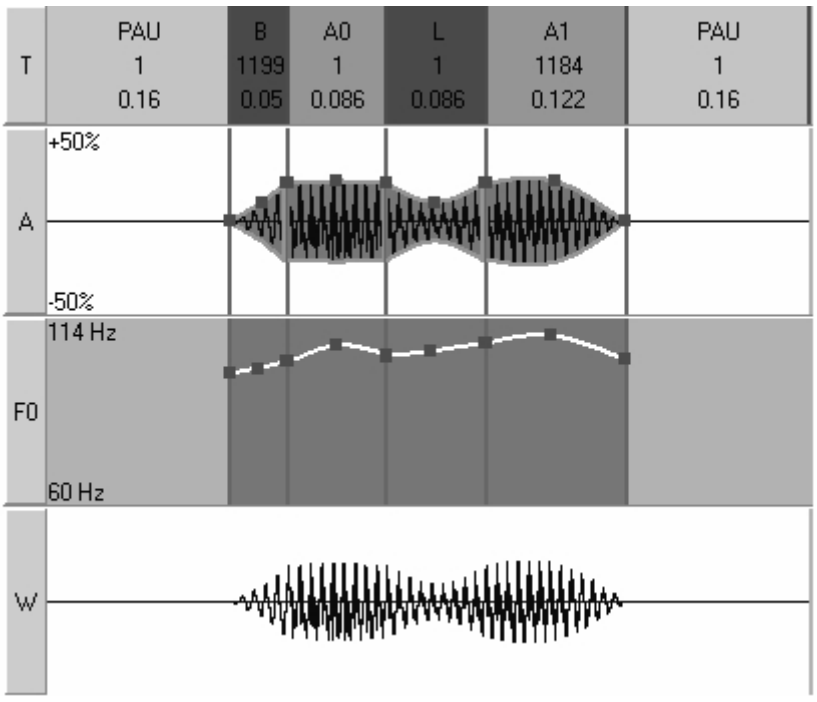

Рис. 1 – Результат компиляции UPL-представления в системе синтеза речи.

После разработки и программной реализации системы следует этап формирования акустической базы данных. На данном этапе необходимо производить оценку качества результатов синтеза и при необходимости вносить соответствующие изменения в систему на уровне баз данных либо алгоритмов. Под «качеством» понимается максимальная разборчивость и естественность звучания синтезированной речи. В данной статье

рассматриваются использованный метод и результаты оценки разборчивости синтезированного речевого сигнала.

Существуют различные методики, которые позволяют оценить разборчивость синтезированного речевого сигнала. Одна из таких методик предлагается в ГОСТ Р 50840-95 «Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости» [6]. Данный стандарт распространяется на тракты (аппаратуру) телефонной проводной и радиосвязи, в которых используется аналоговый речевой сигнал, а также на устройства, содержащие преобразователи речевого сигнала в цифровую форму и на синтезаторы речи. Стандарт устанавливает нормы качества передачи (воспроизведения) речи и среди прочего регламентирует методы измерения и оценки разборчивости речи. Пример проведения подобных измерений и оценки можно найти в [7].

Разборчивость речи определяется как относительное количество (в процентах) правильно принятых элементов (слогов, слов, фраз) артикуляционных единиц.

Разработано специализированное ПО, позволяющее автоматизировать процесс оценки разборчивости синтезированного речевого сигнала. Данное ПО использует в своём составе реализованную систему синтеза казахской речи по фонемному тексту. Сформированы аналогичные предложенным в ГОСТ Р 50840-95 слоговые таблицы с ориентацией на казахский язык.

Задаются следующие основные общие требования и подготовка к измерениям:

1. Измерение проводит бригада аудиторов, не имеющих явных дефектов речи и слуха;
2. Измерения разборчивости речи проводит бригада в возрасте от 18 до 30 лет;
3. Бригада аудиторов проходит предварительное обучение.

В процессе теста-оценки запущенная программа воспроизводит синтезированные слоги в следующем ритме: 1 слог в $(3 \pm 0{,}3)$ с. Аудитор набирает услышанные слоги в помощью клавиатуры на специальной экранной форме. Затем для каждого измерения вычисляется среднее значение разборчивости (S) по формуле:

$$S = \frac{1}{N}\sum_{i=1}^{N} S_i ,$$

где S_i – результат единичного измерения, % (диктор – таблица – аудитор); N – число единичных измерений. Для исключения сомнительных результатов измерений разборчивости вычисляется среднее квадратическое отклонение (СКО) σ по формуле:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(S_i - S)^2}{N-1}} .$$

При этом единичные измерения S_i , для которых $|S_i - S| > 2\sigma$, исключаются и производится вычисление нового среднего значения по формуле:

$$S = \frac{1}{N-k}\sum_{i=1}^{N-k} S_i ,$$

где N – число единичных измерений; k – число исключённых измерений.

В таблице 5 представлены результаты вычислений значений S для синтезированного казахского голоса. За неимением аналогичного синтезатора казахской речи сравнение проводилось относительно разборчивости оригинальной речи казахского диктора. Синтезированная казахская речь имела аналогичную расстановку длительностей пауз и фонем, амплитудные и частотные огибающие. Все речевые сигналы имели сходные

характеристики по частоте дискретизации и разрядности оцифровки. Для оценки казахской речи привлекались аудиторы свободно владеющие казахским языком. Для сравнения в таблице 1 приведены также результаты, которые полученные в [7].

Таблица 1

Результаты оценки разборчивости синтезированного сигнала

Название БД или системы синтеза	Слоговая разборчивость, %	Класс качества
Синтезатор казахской речи	93	Высший
Казахский диктор	96	Высший
Мультифон (БД-М)	91	Высший
Мультифон (БД-Ж)	78	1-й
«Nuance»	55	2-й

По приведённым результатам видно, что в процессе проведённого комплекса работ по оценке и повышению разборчивости синтезированной казахской речи удалось добиться «высшего» класса качества синтеза в соответствии с выбранной методикой.

Таким образом, разработанный язык UPL позволяет задавать и описывать разнообразие фонетических и интонационных форм устной речи. Все исходные данные описываются с помощью унифицированного языкового представления, что позволяет осуществлять гибкое межсистемное взаимодействие и на качественном уровне решать задачи синтеза и распознавания речевого сигнала.

Литература

1. Хокинс Дж., Блейксли С.. Об интеллекте.-М.: Вильямс, 2007. – 240 с.
2. Вартанян И. А. Звук – слух – мозг.–Л.: Наука, 1981. – 176 с.
3. Фостер Дж. Автоматический синтаксический анализ.-М.: Мир, 1974. – 71 с.
4. Амиргалиев Е.Н., Мусабаев Р.Р. Вопросы разработки информационной системы синтеза и распознавания казахской речи. Вестник КазНТУ. – 2008. – № 6/1(70). Ст. 28-34.
5. Амиргалиев Е. Н., Мусабаев Р. Р. Методы анализа и проектирования системы синтеза искусственной речи // Таврический вестник информатики и математики.– 2008.– № 1`2008. С. 51-58.
6. ГОСТ Р 50840-95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. – Введ. 21.11.95. – М.: Госстандарт России: Изд-во стандартов, 1995. – 229 с.
7. Лобанов Б. М., Цирульник Л. И. Компьютерный синтез и клонирование речи. Минск, «Белорусская наука», 2008. – 344 с.