Синтез интонационной составляющей речевого сигнала с применением сплайновой интерполяции

М.Н. Калимолдаев
Е.Н. Амиргалиев
Р.Р. Мусабаев
ДГП „Институт проблем ниформатики и управления“ КН МОН РК
e-mail: rmusab@gmail.com

В данной статье дается описание метода синтеза интонационной составляющей речевого сигнала на основе сплайнов - математически рассчитанных кривых, плавно соединяющих отдельные опорные точки интонационного контура. Данный метод был использован при реализации системы компилятивного синтеза речевого сигнала разрабатываемой в ИПИУ МОН РК. В статье описывается специализированный язык, с помощью которого производиться предварительное описание фонетических и интонационных свойств синтезируемого речевого сигнала. Также приводится описание алгоритмов используемых в процессе расчета гладких параметрических кривых задающих динамику изменения регулируемых параметров. Произведено сравнение предложенного в данной работе метода с методом линейной интерполяции, который используется в большинстве существующих систем синтеза речи. Оценка производилась по критерию минимума суммы квадратов невязок между расчетными значениями по двум методам и натуральным эталонным контуром. В результате для метода линейной интерполяции критерий в среднем равен 0.25, в то время как для предложенного метода значение критерия составляет в среднем 0.07.

Введение

Важной составной частью систем синтеза речевого сигнала по тексту является модуль синтеза интонации. Данный модуль предназначен для генерации интонационного контура и его последующего наложения на синтезируемый сигнал. Естественность синтезированного сигнала в значительной степени определяется качеством задания интонационного контура. В процессе синтеза речи важно осуществлять плавное изменение параметров речевого сигнала. В противном случае синтезированная речь будет обладать неестественным звучанием. Таким образом, при построении систем синтеза и распознавания речи актуальной является задача моделирования плавных речевых интонационных процессов. В данной статье дается описание метода синтеза интонационной составляющей речевого сигнала на основе сплайнов – математически рассчитанных кривых, плавно соединяющих отдельные опорные точки интонационного контура. Данный метод был использован при реализации системы компилятивного синтеза речевого сигнала разрабатываемой в ИПИУ МОН РК.

Известны классические работы ряда зарубежных учёных: Г. Фанта [1], Дж. Флангана [2], С. Фуруи [3], П. Тэйлора [4], Х. Хуанга [5]. Подобные вопросы также изучаются в работах белорусских и российских учёных: Б. М. Лобанова [6], В. Н. Сорокина [7] и др.

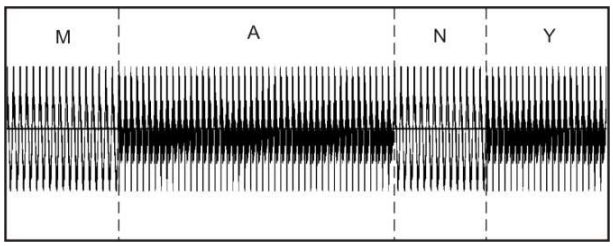

Рис. 1. Исходный речевой сигнал после согласования и конкатенации

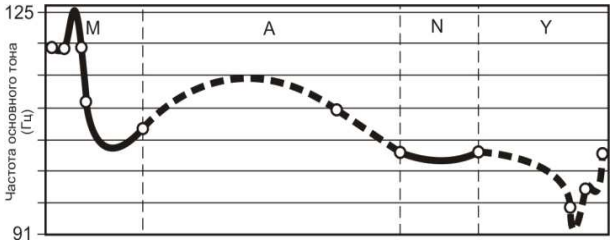

Рис. 2. Наложение контура частоты основного тона

1. Постановка задачи

Для синтеза речевого сигнала по компилятивному принципу необходимо предварительно получить формализованное описание его фонетических и интонационных свойств. В рамках данного описания для всех фонем необходимо указать интонационные характеристики. В их число входит и множество опорных точек параметрических кривых. При этом параметры соседних фонем должны быть плавно согласованны. Таким образом, в качестве задачи ставится разработка специализированного языка, с помощью которого будет производиться предварительное описание фонетических и интонационных свойств синтезируемого речевого сигнала. Также необходимо осуществить алгоритмизацию процесса расчета гладких параметрических кривых, с помощью которых будет задаваться динамика изменения регулируемых параметров.

2. Предложенное решение

На рисунках с 1 по 3 показаны основные этапы синтеза речевого сигнала по компилятивному принципу с применением гладких параметрических кривых заданных ограниченным множеством опорных точек. На рисунке 4 показан результат синтеза.

Для достижения качественного синтеза важно плавно регулировать следующие па-

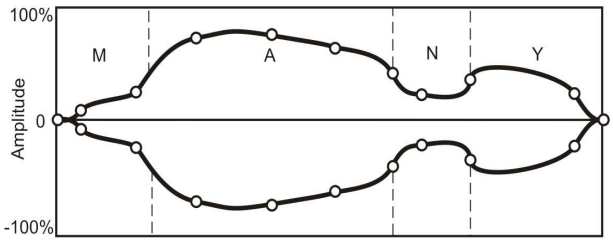

Рис. 3. Наложение амплитудных огибающих

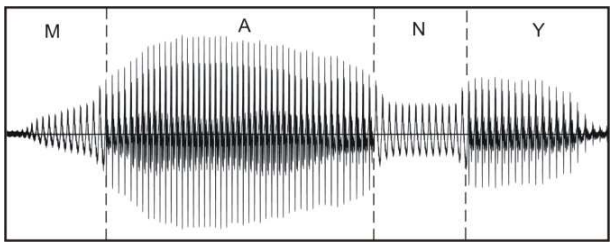

Рис. 4. Результат синтеза

раметры речевого сигнала:

1. Контур частоты основного тона – это главная интонационная составляющая речи (рисунок 2).

2. Амплитудные огибающие, основным назначением которых является динамическое регулирование амплитудного уровня сигнала (рисунок 4). Совместное увеличение амплитуды и частоты сигнала приводит к увеличению его громкости.

При компилятивном синтезе [6] на основе базовых фрагментов речи методом различных алгоритмических манипуляций звуковому сигналу придают необходимую форму. Заданная форма речевого сигнала может зависеть от множества различных факторов: от языка, индивидуальных особенностей голоса, синтезируемого текста, требуемой интонации, скорости и громкости произношения и т. д.

Заранее подготовленный, нормализированный по длительности фонем, общему уровню амплитуд и плавно соединённый из различных фрагментов речевой сигнал подаётся на вход системы регулирования параметров (рисунок 1). В зависимости от требуемых интонационных характеристик формируется контур частоты основного тона и накладывается на исходный речевой сигнал (рисунок 2). Затем на сигнал накладываются амплитудные огибающие (рисунок 3).

Для задания кривой выделяется ограниченное множество опорных точек. Выбирается их оптимальное расположение так чтобы наилучшим образом аппроксимировать исходную функцию контролируемого параметра. Изначально в качестве опорных точек выбираются экстремумы аппроксимируемой функции. В рамках решения задачи выбора оптимального расположения опорных точек требовалось оценить значения их координат в заданных диапазонах поисковым методом в смысле минимума критерия (суммы квадратов невязок). Критерий имеет следующий вид (1):

$$K = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\frac{Y_j^* - Y_j}{Y_j^*}\right)^2}, \qquad (1)$$

где *Yj*, Yj* - соответственно значения аппроксимируемой функции и полученные значения при расчете кривой; N – количество выборок.

Ниже приводится алгоритм вычисления произвольной точки гладкой параметрической кривой.

Входные данные:

1. A – множество опорных точек заданных своими координатами *(X;Y)*

2. Ax – компонента X элемента множества A

3. Ay – компонента Y элемента множества A
4. T – задаёт положение вычисляемой точки на кривой, $t \in [0, 1]$

Выходные данные:

1. X – координата вычисленной точки по оси X
2. Y - координата вычисленной точки по оси Y

Нотация:

1. $f(g) = g^3 - g$
2. i, j – переменные для счётчиков циклов
3. Num – количество элементов множества A
4. dT – значение приращения для T на каждый элемент множества A
5. dX – значение приращения по оси X
6. Px, Py, Wx, Wy, D - множества с количеством элементов равным Num
7. div – операция целочисленного деления

1. Инициализация: Num = *Длина(A) - 1,*
2. Цикл для каждого i = *1..Num-1*
   $D_i = 4$
   $W_x = 6 \cdot ((Ax_{i+1} - Ax_i) - (Ax_i - Ax_{i-1}))$;
   $W_y = 6 \cdot ((Ay_{i+1} - Ay_i) - (Ay_i - Ay_{i-1}))$
   Конец цикла
3. $Px_0 = 0, Py_0 = 0, Px_{Num} = 0, Py_{Num} = 0$
4. Цикл для каждого i = *1..Num-2*
   $Wy_{i+1} = Wy_{i+1} - Wy_i \cdot 0.25$;
   $Wx_{i+1} = Wx_{i+1} - Wx_i \cdot 0.25$
   $D_{i+1} = D_{i+1} - 0.25$
   Конец цикла
5. Цикл для каждого i = *Num-1..1*
   $Px_i = \frac{Wx_i - Px_{i+1}}{D_i}$; $Py_i = \frac{Wy_i - Py_{i+1}}{D_i}$
   Конец цикла
6. $X = Ax_0$; $Y = Ay_0$
7. $dX = Ax_{Num} - Ax_0$ 8. Если $dX > 0$ тогда
   Начало
   $dT = \frac{1}{Num}$
   Цикл для каждого i = 0..Num-1
   Если $(dT \cdot i \leq T)$ и $(dT \cdot (i + 1) \geq T)$ тогда
   Прерывание цикла
   $T = \left(T - \left(T \quad div \quad dT \right) \cdot dT\right) \cdot Num$
   $X = T \cdot Ax_{i+1} + (1 - T) \cdot Ax_i + \frac{f(T) \cdot Px_{i+1} + f(1-T) \cdot Px_i}{6}$

$$Y = T \cdot Ay_{i+1} + (1 - T) \cdot Ay_i + \frac{f(T) \cdot Py_{i+1} + f(1-T) \cdot Py_i}{6}$$

Конец.

Для решения задачи синтеза речевого сигнала [4] используется унифицированный язык фонетического представления (Unified Phonetic Language - UPL). Фактически данный язык является расширенной фонетической транскрипцией. На языке UPL описываются требуемые характеристики речевого сигнала, на основе которых компилятор выбирает наиболее подходящие элементы компиляции и осуществляет последующую генерацию речевого сигнала.

3. Выводы

Произведено сравнение предложенного в данной работе метода с методом линейной интерполяции, который используется в большинстве существующих систем синтеза речи [6]. Оценка производилась по критерию минимума суммы квадратов невязок по аналогии с формулой (1) между расчетными значениями по двум методам и натуральным эталонным контуром. В результате для метода линейной интерполяции критерий в среднем равен 0.25, в то время как для предложенного метода значение критерия составляет в среднем 0.07. Таким образом, параметрическое описание синтезируемого речевого сигнала на языке UPL в совокупности с методом аппроксимации его интонационной составляющей сплайнами позволяют добиться улучшения качества аппроксимации по сравнению с существующими методами.

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 задачу синтеза речевого сигнала.

Предложенный подход может также использоваться в системах распознавания речи и идентификации диктора по особенностям его интонации.

Список литературы

[1] Fant G. Speech Acoustics and Phonetics. Kluwer Academic Publishers, Dordrecht, 2004. 333 pp.

[2] Flanagan J. L., Speech analysis, synthesis and perception. Springer-Verlag, 1972.

[3] Furui S., Digital Speech Processing, Synthesis and Reccognition. Marcel Dekker, 2001.

[4] Taylor P. Text to Speech Synthesis. - University of Cambridge, 2007. 597 pp.

[5] Huang X., Acero A., Hon H.-W. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall, 2001. 472 pp.

[6] Лобанов Б. М., Цирульник Л. И. Компьютерный синтез и клонирование речи. Минск, «Белорусская наука», 2008. – 344 с.

[7] Сорокин В.Н. Синтез речи. Москва, Наука, 1992. 392 с.