Один метод модуляции речевого сигнала по амплитуде и его применение в системах синтеза и клонирования речи

Е. Н. Амиргалиев, Р. Р. Мусабаев
КазНТУ им. К.И. Сатпаева, Алматы, Казахстан
e-mail: amir_ed@mail.ru, rmusab@gmail.com

Рассматривается метод модуляции речевого сигнала по амплитуде. Основное назначение метода — модификация интонационных характеристик речевого сигнала.

Ключевые слова: синтез речи, клонирование речи, речевой сигнал, TTS, интонация, просодия, преобразование текста в речь.

Введение

Существует задача синтеза речевого сигнала с изменяющейся интонацией. Эта задача наиболее часто решается в рамках систем речевого синтеза по тексту, когда на вход системы подается произвольная текстовая информация, а на выходе получается соответствующий речевой сигнал, максимально приближенный к естественной человеческой речи. Имеются также ряд задач по клонированию речевого сигнала, в рамках которых не просто синтезируют качественный речевой сигнал, но стремятся придать ему максимальное сходство с персональными характеристиками речи [1]. Эта технология является технологией двойного назначения.

По данной проблеме известны классические работы ряда зарубежных ученых: Г. Фанта [2], Дж. Фланагана [3], С. Фуруи [4], П. Тэйлора [5], Х. Хуанга [6]. Подобные вопросы изучаются также в работах белорусских и российских ученых: Б.М. Лобанова [1], М.А. Сапожкова [7] и др.

1. Предлагаемый метод

В случае компилятивного синтеза речи в системе присутствует конечное множество базовых фрагментов речевого сигнала: $F = \{f_1; f_2; ...; f_n\}$, где n — общее количество фрагментов. Эти фрагменты получаются в процессе записи речи диктора и последующего автоматического либо ручного их выделения специалистами по фонетике [8]. Размерность базового фрагмента и их количество зависят от выбранного подхода. Наиболее часто используются речевые фрагменты следующих размерностей:

1) полуфон — половина фонемы;
2) фонема — целая элементарная единица;

3) дифон — два смежных полуфона различных фонем с переходной областью между ними;

4) слоги, слова, фразы и т. д.

Общее количество выделенных звуковых фрагментов в системе может колебаться от нескольких сотен до нескольких десятков тысяч. Для повышения качества синтеза достаточно увеличивать количество используемых базовых фрагментов, что приводит к увеличению используемых ресурсов и времени синтеза.

В компилятивной системе речевого синтеза одновременно используются различные типы базовых фрагментов, составляющие конечное множество типов: $T = \{t_1; t_2; ...; t_n\}$, где n — общее число используемых типов. Например, можно выделить следующие типы базовых фрагментов $T = \{V; N; E; P\}$: V — вокализированные, N — шумовые, E — взрывные и щелкающие, P — паузы. Каждому из данных типов соответствует множество объединенных под ним звуковых фрагментов.

Для каждого типа базовых фрагментов устанавливается свой набор правил модификации его интонационных характеристик $R = \{r_1; r_2; ...; r_n\}$, а также множество методов модификации $M = \{m_1(p_{11}; p_{12}; ..; p_{1k}); m_2(p_{21}; p_{22}; ...; p_{2l}); ..; m_n(p_{n1}; p_{n2}; ...; p_{nj})\}$, которыми оперируют данные правила. Каждое правило оперирует одним либо несколькими методами с заданным набором параметров $\{p_{11}; p_{12}; ..; p_{1k}\}$. Правила оперируют также множеством характеристик $C = \left\{\left\{c_1^B; c_2^B; ...; c_n^B\right\}; \left\{c_1^E; c_2^E; ...; c_k^E\right\}\right\}$ как самого базового фрагмента c_n^B, так и его контекстного окружения c_k^E. Различным комбинациям данных характеристик могут быть сопоставлены различные методы интонационной модификации. В общем случае при реализации системы синтеза речи по компилятивному принципу необходимо оперировать следующим комплексным множеством:

$$X = \left(\{F_1; T_1; R_1; M_1; C_1\}; \{F_2; T_2; R_2; M_2; C_2\}; ...; \{F_n; T_n; R_n; M_n; C_n\}\right).$$

Как известно [1], модулирование интонации производится методом изменения длительностей и частотных характеристих различных фрагментов речевого сигнала (в основном это фонемы), а также расстановкой пауз между фонемами. В речевом сигнале наибольшую интонационную составляющую имеют вокализированные участки, что обусловливает особую значимость регулирования их длительностей и частотных характеристик. Для таких типов речевых фрагментов как шумовые участки и паузы можно ограничиться регулированием лишь их длительностей без особого ущерба для общего качества синтеза. Таким образом, для проведения качественного синтеза необходимо оперировать набором методов модификации следующих параметров речевого сигнала:

1) контура частоты основного тона [9];

2) длительностей фонем [10];

3) амплитудной огибающей.

В данной статье предлагается подход для осуществления модификации амплитудной огибающей вокализированных составляющих речевого сигнала. Данный подход был апробирован и успешно применяется в одной из существующих систем синтеза и клонирования речи [11]. Для использования данного метода необходимо предварительно произвести разметку речевого сигнала по частоте основного тона ($F0$) для элементов множества $F \in V$. В результате получаем множество сегментов $S = ((i_1; k_1); (i_2; k_2); ...; (i_n; k_n))$, которые задаются индексом начальной выборки i_n и количеством входящих выборок (рис. 1).

После разметки производится нормализация множества сегментов S по амплитуде. Для этого используются i_n- и i_{n+1}-индексы граничных выборок нормализируемого

микросегмента. Форма сигнала изменяется таким образом, чтобы выровнять выборку с индексом i_{n+1} до уровня выборки i_n. Новое значение амплитудного уровня Z_x для каждой выборки с индексом $i_x \in [i_n; i_{n+1}]$ вычисляется следующим образом:

$$Z_x = Z_x \left(1 + x \frac{1}{i_{n+1} - i_n} \left[\frac{Z_n}{Z_{n+1}} - 1\right]\right),$$

где Z_x — значение амплитудного уровня для рассматриваемой выборки, $x \in [0; i_{n+1} - i_n]$, Z_n и Z_{n+1} — соответственно значения дискретных выборок сигнала с индексами i_n и i_{n+1}, $i_{n+1} - i_n > 0$, $Z_{n+1} \neq 0$. Затем граничные выборки приводятся к заданному амплитудному уровню L, а промежуточные также пропорционально увеличиваются:

$$Z_x = \begin{cases} \text{если} \quad Z_n \neq 0, \quad \text{то} \quad Z_x = Z_x \dfrac{L}{Z_n}, \\ \text{иначе} \qquad\qquad\quad Z_x = Z_x. \end{cases}$$

На рис. 2 проиллюстрирован процесс нормализации сигнала по амплитудному уровню, в итоге которого $h_1 = |h_2| = h_3 = |h_4| = L$. Амплитудная нормализация сигнала позволяет впоследствие применить к нему произвольную огибающую амплитудного уровня и таким образом производить модуляцию сигнала по громкости. Для задания плавных огибающих используются параметрические кривые Безье [12]. С помощью кривой Безье можно аппроксимировать сложные непрерывные формы колебаний, задав всего несколько опорных (характерных) точек, через которые должна пройти данная кривая.

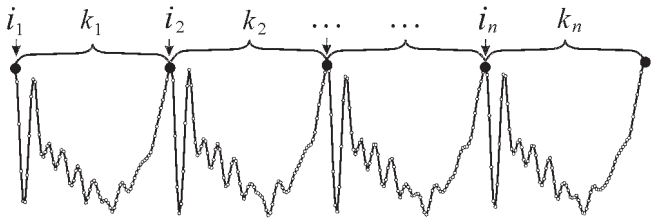

Рис. 1. Исходное сегментированное множество выборок речевого сигнала

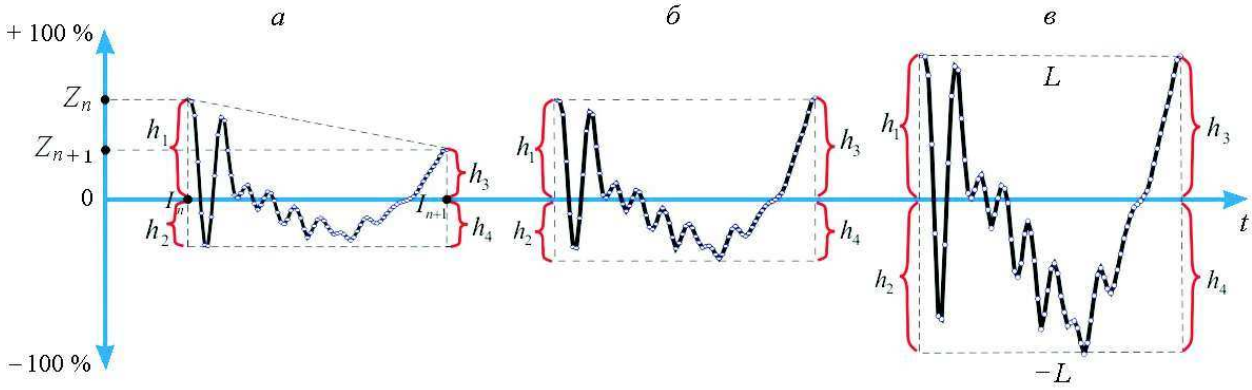

Рис. 2. Процесс нормализации вокализированного микросегмента речевого сигнала по амплитудному уровню: *а* — исходный микросегмент, *б* — нормализация граничных уровней, *в* — приведение общего уровня к заданному

С возрастанием сложности форм аппроксимируемых колебаний необходимо увеличивать количество опорных точек. Кривая Безье задается выражением: $B(t) = \sum_{i=0}^{n} P_i b_{i,n}(t)$, $0 < t < 1$, где P_i является функцией компонент векторов для опорных точек, $b_{i,n}(t)$ — базисные функции кривой Безье (полиномы Бернштейна):

$$b_{i,n}(t) = \left(\begin{array}{c} n \\ i \end{array} \right) t^i (1-t)^{n-i}, \quad \left(\begin{array}{c} n \\ i \end{array} \right) = \frac{n!}{i!\,(n-i)!},$$

здесь n — степень полинома, i — порядковый номер опорной точки. С помощью параметра t определяется точка, принадлежащая кривой. При этом за единицу принимается вся протяженность кривой от начальной до конечной точки.

Координаты (X, Y) произвольной точки, заданной параметром $0 < t < 1$, вычисляются следующим образом:

$$X = T \cdot A_{i+1}^X + (1-T) \cdot A_i^X + \frac{1}{6}\left[f(T) \cdot X_{i+1}^P + f(1-T) \cdot X_i^P\right],$$

$$Y = T \cdot A_{i+1}^Y + (1-T) \cdot A_i^Y + \frac{1}{6}\left[f(T) \cdot Y_{i+1}^P + f(1-T) \cdot Y_i^P\right],$$

где i — индекс ближайшей слева опорной точки из множества $A^{(X,Y)}$, соответствующей условиям $i\dfrac{1}{N_{\max}} \le t$ и $(i+1)\dfrac{1}{N_{\max}} \ge t$; $N_{\max}$ — длина множества $A^{(X,Y)}$ за минусом единицы; A_i^X и A_i^Y — i-й элемент множества $A^{(X,Y)}$, задающий координаты X и Y i-й опорной точки параметрической кривой

$$f(x) = x^3 - x, T = N_{\max}\left(t - D_{\max}\frac{1}{N_{\max}}\right),$$

$$D_{\max} = \begin{cases} \text{если } \; tN_{\max} > 0 \;\text{ и }\; \operatorname{trunc}(tN_{\max}) = 0, & \text{то } D_{\max} = tN_{\max} - 1, \\ \text{иначе} & D_{\max} = trunc\,(tN_{\max}), \end{cases}$$

здесь $\operatorname{trunc}(x)$ — функция округления дробного числа до целой части в меньшую сторону.

Перед непосредственным вычислением координат $(X; Y)$ произвольной точки кривой производится предварительное вычисление величин X_i^P при изменении i в диапазоне $[N_{\max} - 1; 1]$:

$$X_i^P = \frac{1}{D_i}\left(W_i^X - X_{i+1}^P\right), \quad Y_i^P = \frac{1}{D_i}\left(W_i^Y - X_{i+1}^Y\right),$$

где $X_0^P = 0$, $Y_0^P = 0$, $X_{N\max}^P = 0$, $Y_{N\max}^P = 0$. Значения W_i^X, W_i^Y, D_i вычисляются последовательно при изменении i в диапазоне $[1; N_{\max} - 2]$:

$$W_i^X = W_{i+1}^X - \frac{1}{4}W_i^X, \quad W_i^Y = W_{i+1}^Y - \frac{1}{4}W_i^Y, \quad D_{i+1} = D_{i+1} - \frac{1}{4}.$$

При этом их начальные значения задаются при изменении i в диапазоне $[1; N_{\max} - 1]$:

$$W_i^X = 6\left(\left(A_{i+1}^X - A_i^X\right) - \left(A_i^X - A_{i-1}^X\right)\right), \quad W_i^Y = 6\left(\left(A_{i+1}^Y - A_i^Y\right) - \left(A_i^Y - A_{i-1}^Y\right)\right), \quad D_i = 4.$$

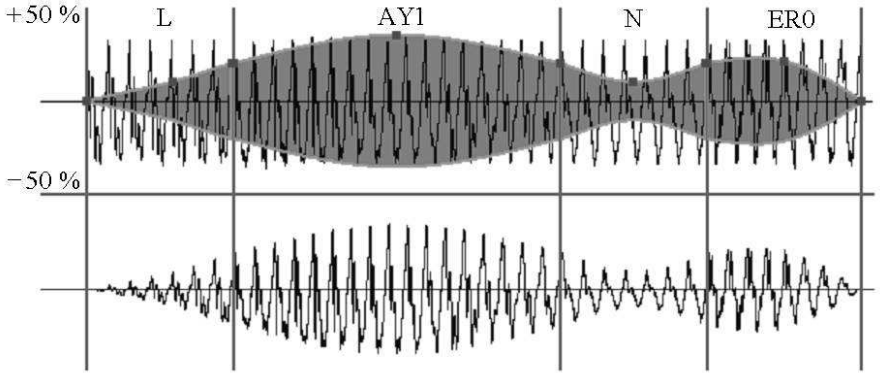

Рис. 3. Процесс модификации амплитуды исходного речевого сигнала по огибающей, заданной набором параметрических кривых Безье

Множества X^P, Y^P, W^Y, W^X и D имеют размерность, равную размерности множества $A^{(X,Y)}$.

Таким образом, имея множество нормализированных дискретных выборок речевого сигнала $Z = \{z_0; z_1; ...; z_{n-1}\}$, где n — количество выборок, а также функцию Безье $\mathrm{Bz}\left(A^{(X,Y)}; t\right)$, которая задается множеством опорных точек

$$A^{(X,Y)} = \left\{ \left(A_1^X; A_1^Y\right); \quad \left(A_2^X; A_2^Y\right); \quad ...; \quad \left(A_m^X; A_m^Y\right) \right\},$$

где m — количество опорных точек, можно осуществить амплитудную модуляцию сигнала, представленного множеством Z:

$$z_i = z_i \cdot \mathrm{Bz}\left(A^{(X,Y)}; t\right), \quad t = \frac{1}{L-1}\left(\frac{i - I_1}{I_2 - I_1} + N_1\right),$$

где L — общее количество опорных точек, I_1 и I_2 — индексы дискретных выборок, соответствующие ближайшей левой и правой опорным точкам для выборки z_i, $I_1 \in [0; n-1]$, $I_2 \in [0; n-1]$, N_1 — номер ближайшей слева опорной точки для выборки z_i, $N_1 \in [0; N_{\max}]$.

На рис. 3 проиллюстрирован процесс модификации амплитуды исходного речевого сигнала по огибающей, заданной набором параметрических кривых Безье. Здесь для каждой фонемы (L, AY1, N, ER0) задается собственная амплитудная огибающая. При этом комплексная огибающая плавно задается общим множеством огибающих каждой из фонем. В приведенном примере

$$A = \left\{ \begin{array}{l} A^L = \left\{ (0;0) \quad (0.6;0.1) \quad (1;0.2) \right\} \\ A^{AY} = \left\{ (0;0.2) \quad (0.5;0.35) \quad (1;0.2) \right\} \\ A^N = \left\{ (0;0) \quad (0.5;0.1) \quad (1;0.2) \right\} \\ A^{ER} = \left\{ (0;0) \quad (0.5;0.21) \quad (1;0) \right\} \end{array} \right\}.$$

Заключение

У рассмотренного метода имеются аналоги. Наиболее часто в компилятивных системах синтеза и клонирования речи установка амплитуд фонем осуществляется усилением (ослаблением) сигналов фонем путем умножения всех значений сигнала на единый

Результаты оценки трудоемкости и разборчивости синтезированного сигнала методов амплитудной модуляции

Метод	Трудоемкость	Разборчивость, %
Модуляция кривой Безье	12503	93
Умножение сигнала на коэффициент	1000	87

коэффициент, задаваемый энергетическим портретом [1]. В ходе проведенного сравнительного анализа методов получены результаты, представленные в таблице.

Трудоемкость метода оценивалась количеством элементарных операций на языке высокого уровня, затрачиваемых на обработку 500 дискретных выборок сигнала. Разборчивость результатов синтеза оценивалась по методике, предложенной в ГОСТ Р 50840-95 [13]. Синтез осуществлялся с помощью одного синтезатора, но с использованием различных методов амплитудной модуляции. По результатам оценок видно, что применение предложенного метода позволяет добиться большей разборчивости синтезированного сигнала. При этом затраты вычислительных ресурсов также значительно увеличиваются.

Список литературы

[1] Лобанов Б.М., Цирульник Л.И. Компьютерный синтез и клонирование речи. Минск: Белорусская наука, 2008.

[2] Fant G. Speech Acoustics and Phonetics. Dordrecht: Kluwer Acad. Publ., 2004.

[3] Фланаган Дж. Анализ, синтез и восприятие речи. М., 1968.

[4] Furui S. Digital Speech Processing, Synthesis, and Recognition. N.Y.: Marcel Dekker Inc., 2001.

[5] Taylor P. Text to Speech Synthesis. Univ. of Cambridge, 2007.

[6] Xuedong Huang, Alex Acero, Raj Reddy. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR, 2001.

[7] Сапожков М.А. Речевой сигнал в кибернетике и связи. М., 1968.

[8] Амиргалиев Е.Н., Мусабаев Р.Р. Алгоритмы выделения и классификации фонем в системах синтеза искусственной речи // Проблемы автоматики и управления: Научно-техн. журн. / Национальная академия наук Кыргызской Республики. Бишкек: Илим, 2008. С. 32–35.

[9] Амиргалиев Е.Н., Мусабаев Р.Р. Определение структуры и способов модификации множества эталонных речевых сигналов в системах синтеза речи // Вестник КазНТУ. 2008. № 6/1(70). С. 25–28.

[10] Мусабаев Р.Р. Технологические особенности модуляции продолжительности речевого сигнала в системах синтеза речи // Сб. тр. междунар. науч.-практ. конф. “Современные проблемы математики, информатики и управления”. Алматы, 2008. С. 98–100.

[11] Амиргалиев Е.Н., Мусабаев Р.Р. Вопросы разработки информационной системы синтеза и распознавания казахской речи // Вестник КазНТУ. 2008. № 6/1(70). С. 28–34.

[12] Мусабаев Р.Р. Использование сплайнов при решении задач генерации речевого сигнала // Вестник КазНУ. 2008. № 4(59). С. 173–175.

[13] Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. ГОСТ Р 50840-95. Введ. 21.11.95. М.: Госстандарт России, 1995. 229 с.

Поступила в редакцию 30 марта 2009 г.,
в переработанном виде — 1 сентября 2009 г.