

МЕТОД МОДУЛЯЦИИ РЕЧЕВОГО СИГНАЛА И ЕГО ПРИМЕНЕНИЕ В СИСТЕМАХ РЕЧЕВОЙ ОБРАБОТКИ

М. Н. Калимолдаев, Р. Р. Мусабаев, О. Ж. Мамырбаев

Институт проблем информатики и управления, 480100, Алма-Ата, Казахстан

УДК 519.7

Рассмотрен метод модуляции речевого сигнала по амплитуде, предназначенный для модификации интонационных характеристик речевого сигнала.

Ключевые слова: синтез речи, клонирование речи, речевой сигнал, text-to-speech, интонация, просодия, преобразование текста в речь.

In the given article the modification method of prosodic and intonational characteristics for periodic components of a speech signal is considered.

Key words: speech synthesis, speech cloning, speech signal, intonation, prosody, text-to-speech.

Введение. Существует задача синтеза речевого сигнала с изменяющейся интонацией. Данная 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. Также существует ряд задач по клонированию речевого сигнала, при решении которых синтезируемому качественному речевому сигналу придается максимальное сходство с персональными характеристиками речи [1]. Данная технология является технологией двойного назначения.

Среди работ по данной теме следует отметить работы [1–7] и др.

Предлагаемый метод. В случае компилятивного синтеза речи в системе имеется конечное множество базовых фрагментов речевого сигнала $F = \{f_1, f_2, \dots, f_n\}$, где n — общее количество фрагментов. Данные фрагменты получаются в процессе записи речи диктора и последующего автоматического либо неавтоматического выделения их специалистами по фонетике [8]. Размерность базовых фрагментов 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 и переходная область 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 необходимо увеличивать количество используемых базовых фрагментов, что приводит к увеличению используемых ресурсов, а также времени синтеза.

В компилятивной системе речевого синтеза одновременно используются различные типы базовых фрагментов, составляющие конечное множество $T = \{t_1, t_2, \dots, t_n\}$, где n — общее количество используемых типов. Например, можно выделить следующие типы базовых фрагментов $T = \{V, N, E, P\}$: V — вокализированные, N — шумовые, E — взрывные и

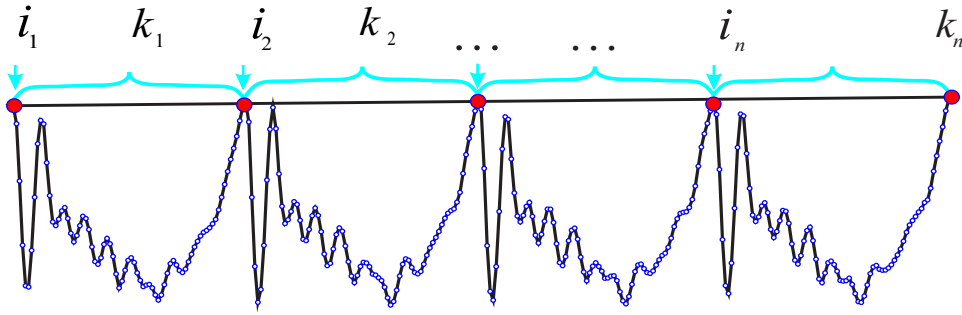


Рис. 1. Исходное сегментированное множество выборок речевого сигнала

щелкающие, P — паузы. Каждому данному типу соответствует множество звуковых фрагментов.

Для каждого типа базовых фрагментов устанавливается набор правил модификации его интонационных характеристик $R = \{r_1, r_2, \dots, r_n\}$, а также множество методов модификации $M = \{m_1(p_{11}, p_{12}, \dots, p_{1k}), m_2(p_{21}, p_{22}, \dots, p_{2l}), \dots, m_n(p_{n1}, p_{n2}, \dots, p_{nj})\}$, которые используются на основании данных правил. Каждое правило оперирует одним либо несколькими методами с заданным набором параметров $\{p_{11}, p_{12}, \dots, p_{1k}\}$. Правила оперируют также множеством характеристик $C = \{\{c_1^B, c_2^B, \dots, c_n^B\}, \{c_1^E, c_2^E, \dots, c_k^E\}\}$ как самого базового фрагмента c_n^B , так и его контекстного окружения c_k^E . Различным комбинациям данных характеристик могут быть поставлены в соответствие различные методы интонационной модификации. В общем случае при реализации системы синтеза речи по компилятивному принципу необходимо оперировать комплексным множеством $X = (\{F_1, T_1, R_1, M_1, C_1\}, \{F_2, T_2, R_2, M_2, C_2\}, \dots, \{F_n, T_n, R_n, M_n, C_n\})$.

Как известно, модулирование интонации производится путем изменения длительностей и частотных характеристик различных фрагментов речевого сигнала, в основном фонем, а также расстановки пауз между фонемами [1]. В речевом сигнале наибольшую интонационную составляющую имеют вокализированные участки, что обуславливает особую значимость регулирования их длительностей и частотных характеристик. Для таких типов речевых фрагментов, как шумовые участки и паузы, без ущерба для качества синтеза можно ограничиться регулированием лишь их длительностей. Таким образом, для осуществления качественного синтеза необходимо использовать набор методов модификации следующих параметров речевого сигнала:

- контура частоты основного тона [9];
- длительностей фонем [10];
- амплитудной огибающей.

В настоящей работе предлагается подход для осуществления модификации амплитудной огибающей вокализированных составляющих речевого сигнала. Данный подход был апробирован и применяется в одной из систем синтеза и клонирования речи [11]. Для того чтобы использовать этот метод, предварительно необходимо выполнить разметку речевого сигнала по частоте основного тона F_0 для элементов множества $F \in V$. В результате получаем множество сегментов $S = ((i_1, k_1), (i_2, k_2), \dots, (i_n, k_n))$, которые задаются индексом начальной выборки i_n и количеством входящих выборок k_n (рис. 1).

После разметки производится нормализация множества сегментов S по амплитуде, для этого используются индексы граничных выборок нормализуемого микросегмента i_n и i_{n+1} . Форма сигнала изменяется таким образом, чтобы выборка с индексом i_{n+1} была выровнена

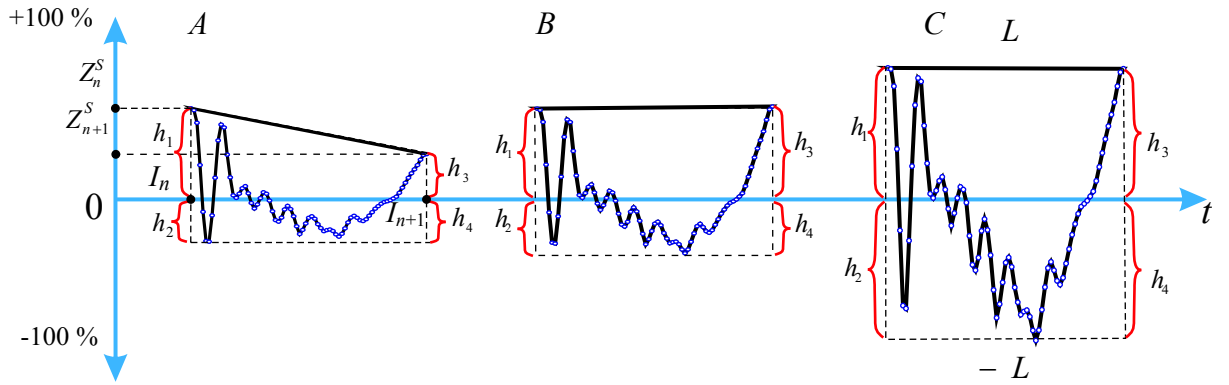


Рис. 2. Процесс нормализации вокализованного микроsegmenta речевого сигнала по амплитудному уровню:

A — исходный микроsegment, *B* — нормализация граничных уровней, *C* — приведение общего уровня к заданному

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:

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

Здесь Z_x — дискретное значение речевого сигнала (выборки) при импульсно-кодовой модуляции, при этом мгновенное значение аналогового сигнала измеряется через равные промежутки времени; $x \in [0, i_{n+1} - i_n]$; Z_n, 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:

$$Z_x = \begin{cases} Z_x \frac{L}{Z_n}, & Z_n \neq 0, \\ Z_x, & Z_n = 0. \end{cases}$$

На рис. 2 представлен процесс нормализации сигнала по амплитудному уровню, в результате которого $h_1 = |h_2| = h_3 = |h_4| = L$. Амплитудная нормализация сигнала позволяет впоследствии применить к нему произвольную огибающую амплитудного уровня и таким образом произвести модуляцию сигнала по громкости. Для задания плавных огибающих используются параметрические кривые Безье [12]. С помощью кривой Безье можно аппроксимировать сложные непрерывные формы колебаний, задав лишь несколько опорных (характерных) точек, через которые должна пройти данная кривая. При увеличении сложности форм аппроксимируемых колебаний достаточно увеличивать количество опорных точек. Кривая Безье задается выражением

$$B(t) = \sum_{i=0}^n P_i b_{i,n}(t), \quad 0 < t < 1,$$

где P_i — функция компонент векторов для опорных точек; $b_{i,n}(t)$ — базисные функции кривой Безье (полиномы Бернштейна):

$$b_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i}, \quad \binom{n}{i} = \frac{n!}{i!(n-i)!},$$

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

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

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

$$Y = TA_{i+1}^Y + (1-T)A_i^Y + \frac{1}{6} [f(T) Y_{i+1}^P + f(1-T) Y_i^P].$$

Здесь i — индекс ближайшей слева опорной точки из множества $A^{(X,Y)}$, соответствующей условиям $i/N_{\max} \leq t$ и $(i+1)/N_{\max} \geq t$; N_{\max} — длина множества $A^{(X,Y)}$ минус единица; A_i^X, A_i^Y — соответственно i -е элементы множества $A^{(X,Y)}$, задающие координаты X и Y i -й опорной точки параметрической кривой;

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

$$D_{\max} = \begin{cases} tN_{\max} - 1 & \text{при } tN_{\max} > 0, \text{ trunc}(tN_{\max}) = 0, \\ \text{trunc}(tN_{\max}), & \text{иначе,} \end{cases}$$

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

Перед непосредственным вычислением координат (X, Y) произвольной точки кривой проводится расчет следующих значений при изменении i в диапазоне $[N_{\max} - 1, 1]$:

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

Здесь $X_0^P = 0$; $Y_0^P = 0$; $X_{N_{\max}}^P = 0$; $Y_{N_{\max}}^P = 0$. Значения величин 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 $[1, N_{\max} - 1]$:

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

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

Таким образом, имея множество нормализованных дискретных выборок речевого сигнала $Z = \{z_0, z_1, \dots, z_{n-1}\}$, где n — количество выборок, а также функцию Безье $Bz(A^{(X,Y)}, t)$, которая задается множеством опорных точек $A^{(X,Y)} = \{(A_1^X, A_1^Y), (A_2^X, A_2^Y), \dots, (A_m^X, A_m^Y)\}$, где m — количество опорных точек, можно осуществить амплитудную модуляцию сигнала, представленного множеством Z :

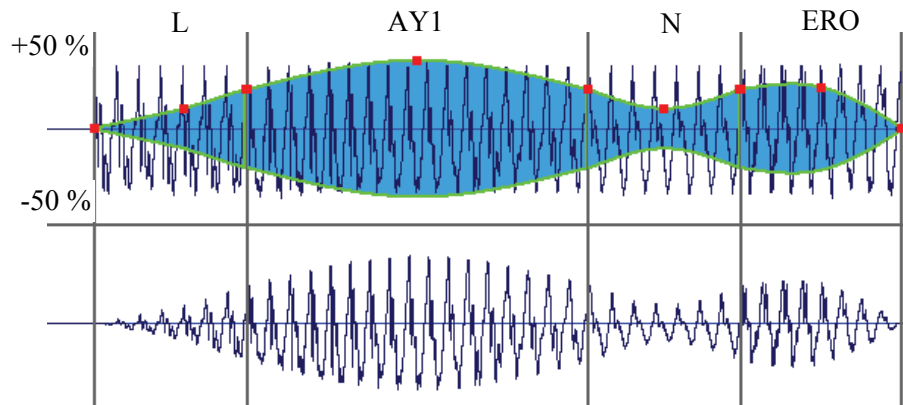


Рис. 3. Процесс модификации амплитуды исходного речевого сигнала по огибающей, заданной набором параметрических кривых Безье

Результаты оценки трудоемкости и разборчивости методов амплитудной модуляции

Метод	Трудоемкость	Разборчивость, %
Модуляция кривой Безье	12 503	93
Умножение сигнала на коэффициент	1000	87

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

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

На рис. 3 представлен процесс модификации амплитуды исходного речевого сигнала по огибающей, заданной набором параметрических кривых Безье. Для каждой фонемы (L, AY, N, ER) задается собственная амплитудная огибающая. При этом комплексная огибающая плавно задается общим множеством огибающих каждой фонемы. В приведенном примере

$$A = \left\{ \begin{array}{l} A^L = \{(0;0) (0,6;0,1) (1;0,2)\} \\ A^{AY} = \{(0;0,2) (0,5;0,35) (1;0,2)\} \\ A^N = \{(0;0) (0,5;0,1) (1;0,2)\} \\ A^{ER} = \{(0;0) (0,5;0,21) (1;0)\} \end{array} \right\}.$$

Заключение. У предлагаемого метода имеются аналоги. Наиболее часто в компилятивных системах синтеза и клонирования речи установка амплитуд фонем осуществляется за счет усиления (ослабления) сигналов фонем путем умножения всех значений сигнала на единый коэффициент, задаваемый энергетическим портретом [1]. В ходе проведенного сравнительного анализа методов получены результаты, представленные в таблице. Трудоемкость метода оценивалась как количество элементарных операций на языке высокого уровня, затрачиваемых на обработку 500 дискретных выборок сигнала. Разборчивость результатов синтеза оценивалась по методике, предложенной ГОСТ Р 50840-95 [13]. Синтез осуществлялся с помощью одного синтезатора, но с использованием различных методов амплитудной

модуляции. Результаты проведенных оценок показывают, что с использованием предложенного метода можно добиться большей разборчивости синтезированного сигнала. При этом затраты вычислительных ресурсов также значительно увеличиваются.

Список литературы

1. ЛОБАНОВ Б. М. Компьютерный синтез и клонирование речи / Б. М. Лобанов, Л. И. Цирульник. Минск: Белорус. наука, 2008.
2. FANT G. Speech acoustics and phonetics. Dordrecht: Kluwer Acad. Publ., 2004.
3. ФЛАНАГАН Дж. Анализ, синтез и восприятие речи. М.: Связь, 1968.
4. FURUI S. Digital speech processing, synthesis, and recognition. N. Y.: Marcel Dekker Inc., 2001.
5. TAYLOR P. Text to speech synthesis. Cambridge: Univ. of Cambridge, 2007.
6. XUEDONG HUANG. Spoken language processing: A guide to theory, algorithm and system development / Xuedong Huang, Alex Acero, Hsiao-Wuen Hon. New Jersey: Prentice Hall, 2001.
7. САПОЖКОВ М. А. Речевой сигнал в кибернетике и связи. М.: Связьиздат, 1968.
8. АМИРГАЛИЕВ Е. Н., МУСАБАЕВ Р. Р. Алгоритмы выделения и классификации фонем в системах синтеза искусственной речи // Пробл. автоматизации и управления (Бишкек). 2008. С. 32–35.
9. АМИРГАЛИЕВ Е. Н., МУСАБАЕВ Р. Р. Определение структуры и способов модификации множества эталонных речевых сигналов в системах синтеза речи // Вестн. КазНТУ. 2008. № 6. С. 25–28.
10. МУСАБАЕВ Р. Р. Технологические особенности модуляции продолжительности речевого сигнала в системах синтеза речи // Тр. Междунар. науч.-практ. конф. “Современные проблемы математики, информатики и управления”, Алма-Ата, 5 нояб. 2008 г. Алма-Ата: Эверо, 2008. С. 98–100.
11. АМИРГАЛИЕВ Е. Н., МУСАБАЕВ Р. Р. Вопросы разработки информационной системы синтеза и распознавания казахской речи // Вестн. КазНТУ. 2008. № 6. С. 28–34.
12. МУСАБАЕВ Р. Р. Использование сплайнов при решении задач генерации речевого сигнала // Вестн. КазНТУ. 2008. № 4. С. 173–175.
13. ГОСТ Р 50840-95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. Введ. 21.11.95. М.: Госстандарт России, 1995. 229 с.

*Калимолдаев Максат Нурадилович — д-р физ.-мат. наук, проф.,
директор Института проблем информатики и управления МОН РК;
тел. +7-727-272-37-11; e-mail: mnk@ipic.kz;*
*Мусабаев Рустам Рафикович — канд. техн. наук,
ученый секретарь Института проблем информатики и управления МОН РК;
e-mail: rmusab@gmail.com;*
*Мамырбаев Оркен Жумажанович — докторант PhD
Института проблем информатики и управления МОН РК;
e-mail: morkenj@mail.ru*

Дата поступления — 18.08.11 г.