УДК 004.934.2

ИСПОЛЬЗОВАНИЕ ПЕРИОДИЧНОСТИ РЕЧЕВОГО СИГНАЛА ПРИ ФОНЕМНОЙ СЕГМЕНТАЦИИ РЕЧИ*

А.Г. ДАВЫДОВ, Б.М. ЛОБАНОВ

Объединенный институт проблем информатики НАН Беларуси
Сурганова, 6, Минск, 220012, Беларусь

Поступила в редакцию 13 апреля 2006

В работе рассматриваются вопросы использования информации о периодичности речевого сигнала при фонемной сегментации речи на основе метода динамического программирования. Предлагается способ разметки речевого сигнала на периоды основного тона, основанный на анализе амплитуды огибающей сигнала и кратковременной функции среднего значения разности, что позволяет совместить положения импульсов возбуждения голосового тракта с границами фонем для уменьшения искажений при синтезе речевого сигнала. Определяются оптимальные значения коэффициентов системы сегментации.

Ключевые слова: анализ речи, период основного тона, динамическое программирование.

Введение

Во многих областях речевых технологий требуются устойчивые методы сегментации речи. Сегментация является важным этапом при начальном обучении системы распознавания речи и при верификации речи диктора [1, 2]. Для систем компиляционного синтеза речи по тексту сегментация записи естественной речи является наиболее трудоемким этапом в процессе создания базы элементов синтеза (аллофонов, дифонов, слогов и т.д.) для нового голоса, в значительной мере определяющим качество результата [3, 4]. Для таких систем большое значение также имеет корректное объединение участков естественной речи, при котором не возникает скачка основного тона в месте объединения. Для выполнения данного условия на этапе создания нового голоса для системы синтеза используется предварительное выравнивание траектории частоты основного тона — т.е. превращение голоса в монотонный. Другим немаловажным аспектом в выполнении указанного условия является совмещение границ вокализованных фонем с импульсами возбуждения основного тона.

Структура системы сегментации речи

Система сегментации речи на основе метода динамического программирования (ДП) и определение оптимальных параметров ее работы подробно рассматривается в работах [5–8]. Обобщенная структура системы приведена на рис. 1. Дальнейшим развитием системы является расширение пространства признаков за счет использования информации о периодичности речевого сигнала (меры тона).

* Данное исследование выполнялось при поддержке европейского фонда INTAS в рамках проекта INTAS № 04-77-7404.

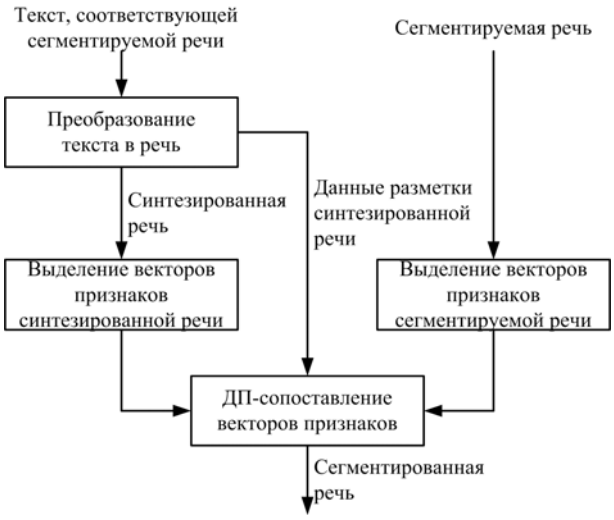

Рис. 1. Структура системы сегментации речи на основе метода динамического программирования

Анализ периодичности речевого сигнала

Один из наиболее эффективных (по критерию вычислительной сложности) методов определения периодичности речевого сигнала базируется на вычислении кратковременной функции среднего значения разности (КФСР) [9]

$$\gamma_n(k) = \frac{1}{M}\sum_{m=0}^{M-1}\left|x(n+m) - x(n+m-k)\right|.$$

Очевидно, что при величине задержки $k = \pm T_0, \pm 2T_0, \ldots$, функция $\gamma_n(k)$ будет иметь глубокие провалы для квазипериодических сигналов, где T_0 — период основного тона, M — интервал интегрирования. Интегрирование целесообразно осуществлять на интервале $M = 1/F_{0\min}$, т.е. таком, что бы КФСР вычислялась хотя бы на одном полном периоде основного тона. Достоинством данного метода определения периодичности речевых сигналов является значительно меньшая вычислительная сложность по сравнению с широко распространенным автокорреляционным методом. Примеры нормированной функции КФСР, вычисленные для вокализованного и невокализованного речевых сигналов, приведены на рис. 2.

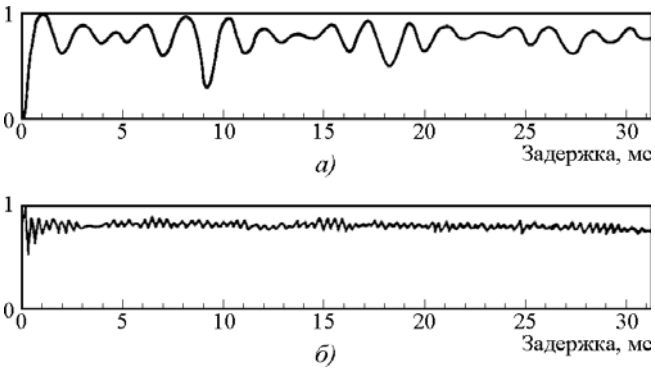

Рис. 2. Нормированная функция КФСР для вокализованного (*а*) и невокализованного (*б*) участков речи

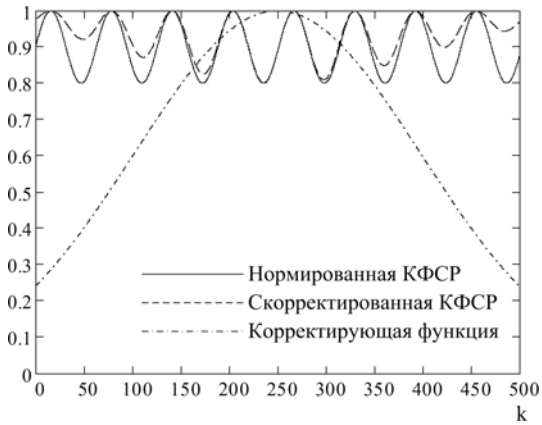

Рис. 3. Пример коррекции нормированной функции КФСР

Предполагая, что частота основного тона располагается в диапазоне от $F_{0\min}$ до $F_{0\max}$ целесообразно пропустить исходный сигнал через полосовой фильтр с полосой пропускания, равной принятому диапазону частоты основного тона.

Меру тона речевого сигнала предлагается оценивать по величине минимума нормированной функции КФСР при значении задержки от $1/F_{0\max}$ до $1/F_{0\min}$:

$$\mu_n = 1 - \min\left(\hat{\gamma}_n(k)\right), \text{ для } k = \left[\frac{1}{F_{0\max}}, \frac{1}{F_{0\min}}\right].$$

Полученные величины меры тона и усредненной конечной разности (КР) меры тона, умноженные на их весовые коэффициенты, включаются в вектор признаков речевого сигнала. В соответствии со структурой системы сегментации речи, представленной в [6], следующим шагом в расстановке границ является выравнивание временных шкал на основе метода динамического программирования.

Своеобразной опциональной постобработкой является совмещение полученных границ фонем с импульсами возбуждения основного тона. Позиции импульсов возбуждения основного тона можно определить исходя из информации об изменениях периода основного тона и огибающей речевого сигнала.

Мгновенное значение периода основного тона можно определить как значение задержки, при котором было определено минимальное значение функции КФСР. Для устранения ошибок неверного определения данной величины, что часто встречается при смене вокализованного и невокализованного участков речи, предлагается использовать процедуру коррекции функции КФСР и переоценки мгновенной величины периода основного тона. Процедура коррекции заключается в нахождении вокализованных регионов, определении наиболее частотного значения периода основного тона для каждого региона и коррекции функции КФСР с учетом полученной статистики.

Вокализованным регионом предлагается считать интервал речевого сигнала заданной минимальной длительности, на котором мера тона не принимает значения меньше определенного порога. Опытным путем было установлено, что минимальная длительность вокализованного региона равна приблизительно трем-четырем периодам основного тона, что для голоса с частотой основного тона 100 Гц соответствует приблизительно 35 мс.

Для исключения ошибок определения периода основного тона предлагается корректировать нормированную КФСР в соответствии со следующими формулами:

$$\tilde{\gamma}_n(k) = \left(\hat{\gamma}_n(k) - 1\right) e^{\alpha(k-\tau)^2} + 1,$$

$$\alpha = \frac{4\ln(s)}{\tau^2},$$

где $\hat{\gamma}_n(k)$ — нормированная КФСР; $\tilde{\gamma}_n(k)$ — скорректированная КФСР; τ — наиболее частотное значение периода основного тона на вокализованном регионе; s — величина от 0 до 1, определяющая крутизну корректирующей функции, как ее значение в точке $\tau/2$. Пример коррекции для функции $\hat{\gamma}_n(k) = 0{,}1\sin(0{,}1\,k) + 0.9$, $\tau = 250$ и $s = 0{,}7$ приведен на рис. 3.

После проведенной коррекции необходимо заново оценить период основного тона.

Для нахождения позиций импульсов возбуждения (питчей) голосового тракта вычисляется огибающая исходного речевого сигнала на основе преобразования Гильберта, подвергнутая низкочастотной фильтрации с частотой среза $F_{0\max}$. Для определения позиции первого питча в вокализованном регионе отыскивается позиция, где мера тона принимает максимальное значение. В окрестности найденной позиции, равной $\pm T_0$, отыскивается позиция минимального значения огибающей, принимаемая как позиция первого питча. Остальные питчи в вокали-

зованном регионе расставляются последовательно в обе стороны от первого питча на основании изменения мгновенного значения периода основного тона (рис. 4).

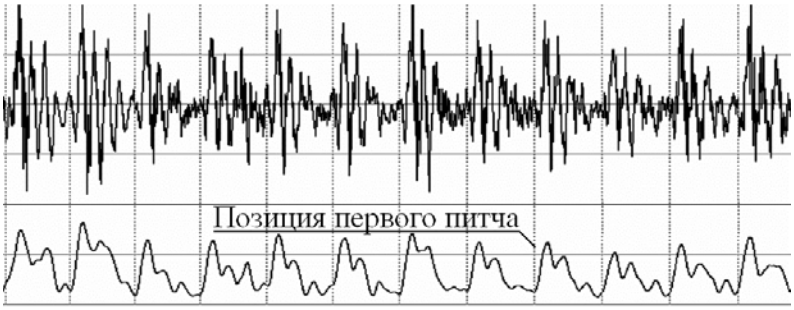

Рис. 4. Разметка вокализованного региона на примере фонемы Λ

Корректировка границ фонем на позиции питчей осуществлялась по критерию ближайшего соседства, при условии, что смещение границы не окажется больше чем 25 мс. Данная величина максимально возможного смещения была определена эмпирическим путем.

Определение оптимальных параметров

Для определения оптимальных параметров сегментации речевого сигнала использовалось тестовое множество (состоящее из 1128 элементов) и целевая функция (равная сумме средней и среднеквадратической ошибки сегментации) из [6]. Граничные значения частоты основного тона принимались равными $F_{0\min} = 50$ и $F_{0\max} = 400$ Гц.

Оптимальные параметры работы системы сегментации речи, определенные в [6], представлены в таблице. Данные параметры использовались как начальные для проведения экспериментов.

Параметры работы системы сегментации речи

Параметр	Значение
Интервал дискретизации сонограммы	1 мс
Интервал усреднения спектра	1 мс
Коэффициенты приведения	0,004
Интервал усреднения КРС	14 мс
Весовой коэффициент спектра	1
Весовой коэффициент КРС	8,1
Коэффициент горизонтального перехода	1
Коэффициент вертикального перехода	1
Коэффициент диагонального перехода	1
Весовой коэффициент времени	1
Ширина начала допустимого интервала	0,05
Ширина конца допустимого интервала	0,35

Очередным шагом в определении оптимальных параметров системы сегментации является оценка коэффициента меры тона, при которой наблюдается минимальное значение целевой функции (рис. 5). Для получения данной оценки анализировался диапазон значений коэффициента меры тона от 0 до 40 включительно с шагом 0,5 при коэффициенте КР меры тона равном 0 и пороге меры тона, равном 1.

Оптимальное значение целевой функции наблюдается при коэффициенте меры тона, равном 19, используемом при дальнейшем определении оптимальных параметров. Для определения значения коэффициента КР меры тона, минимизирующего сумму среднего и среднего квадратического значения ошибки сегментации, анализировался интервал его значений от 0 до 160 с шагом 1 (рис. 6). Как видно из полученной зависимости, наиболее точная сегментация наблюдается при коэффициенте КР меры тона, равного 0, т.е. исключения КР меры тона из вектора признаков.

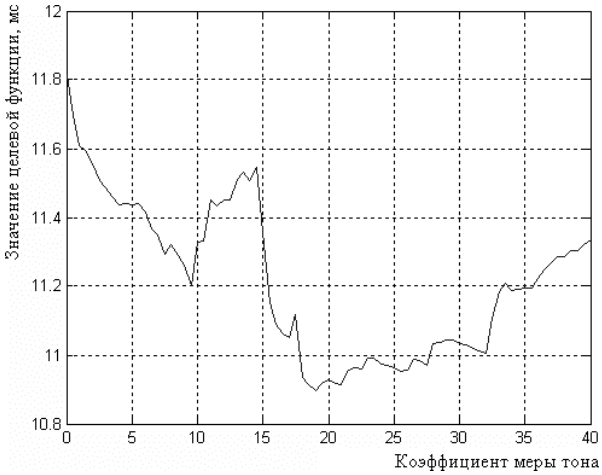

Рис. 5. Определение коэффициента меры тона

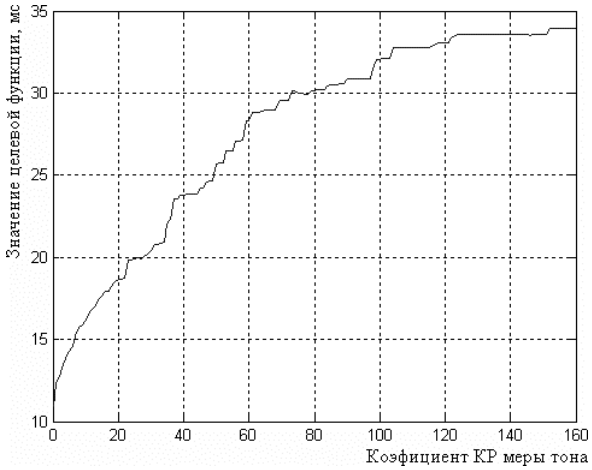

Рис. 6. Определение коэффициента КР меры тона

Рассмотренная выше оптимизация значений коэффициентов работы системы сегментации проводились при пороге меры тона, равном 1, т.е. отсутствии разметки на периоды основного тона. Если же для решения поставленной задачи требуется выполнить такую разметку, то является необходимым оценить и оптимальное значение порога меры тона, при котором отсутствует разметка на невокализованных регионах и полностью промаркированы вокализованные. Для определения такого значения анализировался интервал изменения порога меры тона от 0,1 до 1 с шагом 0,02 (рис. 7).

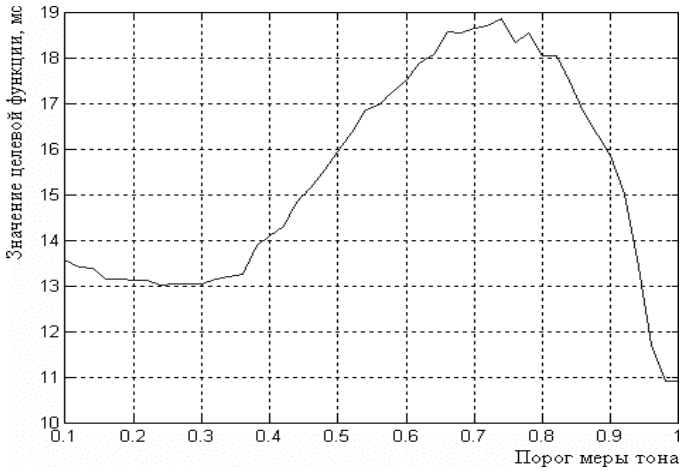

Рис. 7. Определение порога меры тона

Как видно из полученной зависимости, минимальная ошибка достигается при отказе от процедуры совмещения границ фонем с периодами основного тона, а наиболее точная разметка с использованием процедуры выравнивания достигается при пороге меры тона, равном 0,24.

Заключение

Проанализировано использование информации о периодичности речевого сигнала при фонемной сегментации речи. Определены оптимальные параметры работы системы сегментации с учетом квазипериодической структуры вокализованных звуков, позволяющие уменьшить среднюю ошибку сегментации до 4 мс и среднее квадратическое значение ошибки до 6,8 мс при отсутствии выравнивания границ фонем на позиции импульсов возбуждения основного тона и соответственно 5,3 мс и 7,7 мс при использовании выравнивания.

THE SPEECH SIGNAL PERIODICITY UTILIZATION IN PHONEME SPEECH SEGMENTATION

A.G. DAVYDAU, B.M. LOBANOV

Abstract

The problems of the periodicity information utilization during phoneme speech segmentation (on the bases of dynamic programming) are discussed. The method of speech signal marking is suggested. This method is based on the envelope amplitude and short-time mean difference analysis. This allows to align the phoneme boundaries with vocal tract pulses excitation. The optimum speech segmentation system factors are determined.

Литература

1. *Сорокин В.Н., Цыплухин А.И.* // Информационные процессы. 2004. Т. 4, № 2. С. 202–220.
2. *Horak P.* // Improvements in speech synthesis. 2001. P. 331–340.
3. *Лобанов Б.М., Киселев В.В.* // Труды Международной конференции "Диалог-2003". С. 417–424.
4. *Lobanov B.M., Tsirulnik L.I.* // Труды Международной конференции "Речь и компьютер" SPECOM'2004. СПб., 2004. С. 17–21.
5. *Malfrere F., Dutoit T.* // Proc. of Eurospeech'97. 1997. P. 2631–2634.
6. *Давыдов А.Г.* // Информатика. 2006. №1 (9). С. 47–57.
7. *Давыдов А.Г., Киселев В.В., Лобанов Б.М., Цирульник Л.И.* // Изв. Белорус. инж. акад. 2004. № 1/1 С. 112–115.
8. *Sethy A., Narayanan S.* // Proc. of ICSLP 2002–INTERSPEECH. 2002. P. 149–152.
9. *Рабинер Л.Р., Шафер Р.В.* Цифровая обработка речевых сигналов. М., 1981.