УДК 621.391

СЕГМЕНТАЦИЯ РЕЧИ НА ОСНОВЕ МЕТОДА МОДИФИЦИРОВАННОГО НЕПРЕРЫВНОГО ДИНАМИЧЕСКОГО ПРОГРАММИРОВАНИЯ

А.Г. ДАВЫДОВ, Б.М. ЛОБАНОВ

Объединенный институт проблем информатики НАН Беларуси
Сурганова, 6, Минск, 22013, Беларусь

Поступила в редакцию 2 сентября 2004

Рассматривается построение системы сегментации речевого сигнала на основе метода модифицированного непрерывного динамического программирования с использованием системы компиляционного синтеза речи по тексту (TTS).

Ключевые слова: речь, сегментация речи, динамическое программирование.

В последнее время все большее распространение получают системы компиляционного синтеза речи [1], заключающиеся в компиляции участков естественной речи. При этом весьма трудоемким процессом является создание новых баз дикторов для синтеза речи. Основные сложности при создании новой базы связаны с разделением записи голоса диктора на элементы синтеза, которые бы могли в дальнейшем использоваться для синтеза речи новым голосом.

С целью автоматизации сегментирования речевого сигнала на базе метода, описанного в [2], был разработан следующий метод, общая последовательность обработки данных в котором разбивается на следующие этапы:

подготовка эталонных данных при помощи синтезатора речи;

синтез речи по заданному тексту с сохранением структуры (позиций аллофонов в синтезированной речи);

вычисление сонограммы эталонного сигнала;

нормирование сонограммы эталона;

подготовка обрабатываемых данных;

вычисление сонограммы обрабатываемого сигнала;

нормирование сонограммы обрабатываемого сигнала;

нелинейное по времени динамическое сопоставление сонограмм синтезированной и естественной речи;

сегментация данных (перенос меток);

Синтез речи по заданному тексту подробно описан в [1]. Необходимо только заметить, что для сегментации речевого сигнала понадобится не только его синтезированный аналог, но и разметка синтезированной речи на ее элементы.

Вычисление сонограммы может быть выполнено несколькими способами, однако все они преследуют одну цель: выделить в речи такие компоненты, которые бы наиболее хорошо показывали отличие различных звуков друг от друга и кроме этого минимально отличались для различных дикторов. Методы, описанные в [3–5], используют для этого первоначальное вычисление спектра сигнала через преобразование Фурье, а затем приведение его к частотной шкале в барках [3, 4] или в мелах [5]. Для достижения лучшей дикторонезависимости в [3, 4, 5] используется вычисление кепстральных коэффициентов.

В разработанном методе для анализа спектра используется набор из 20 полосовых фильтров Чебышева 3-го порядка, с шириной полосы в 1, 2 или 3 барка. В ходе исследований было установлено, что использование фильтров с полосами в 3 барка дает наименьшую диктор-скую вариативность спектра при незначительном ухудшении отличий одного звука от другого.

Нормирование сонограммы предлагается осуществлять в соответствии со следующими формулами:

$$Sum(n, j) = \frac{1}{T \cdot C - 1} \sum_{k=n-T/2}^{n+T/2} \sum_{l=0}^{C} \Delta(S(n, j), S(k,l)),$$

$$Sn(n, j) = \begin{cases} Sum(n, j), & \text{если} \quad Sum(n, j) \geq 0 \\ 0, & \text{если} \quad Sum(n, j) < 0 \end{cases},$$

где $Sn(n, j)$ — нормированное значение точки n,j сонограммы; $S(n, j)$ — ненормированное значение; T — интервал нормирования; C — число каналов в сонограмме.

Функция $\Delta(S(n, j), S(k,l))$ вычисляется по формуле:

$$\Delta(S(n, j), S(k,l)) = \begin{cases} 1, & \text{если} \quad S(n, j) - S(k,l) > \varepsilon \\ 0, & \text{если} \quad -\varepsilon \leq S(n, j) - S(k,l) \leq \varepsilon \\ -1, & \text{если} \quad S(n, j) - S(k,l) < -\varepsilon \end{cases},$$

где ε — порог шумов.

Примером такого нормирования может служить рис. 1, где на рис. 1,*а* изображены три ненормированные функции, а на рис 1,*б* — они же нормированные.

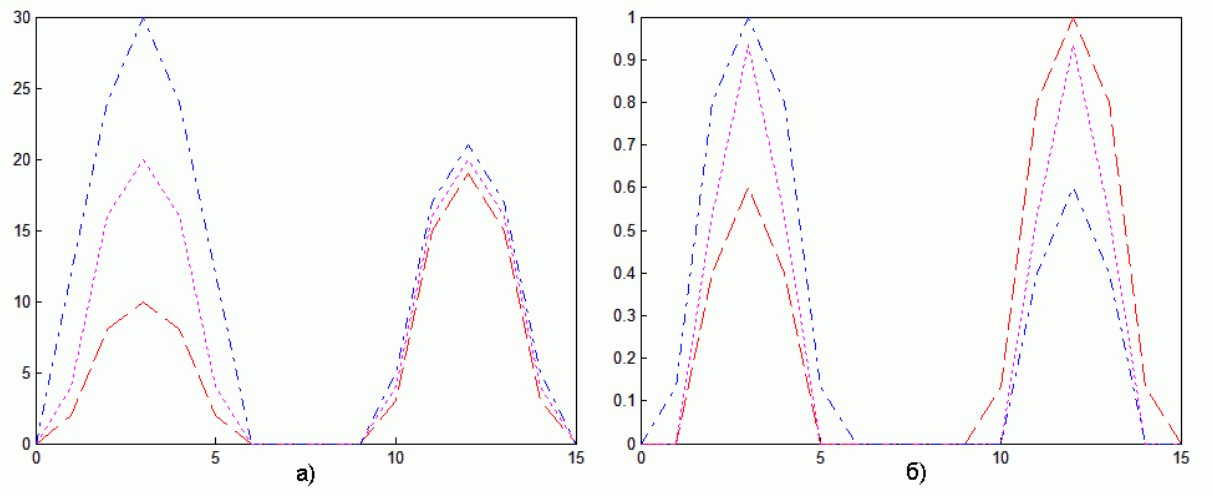

Рис. 1. Пример вычисления нормированного сигнала

Нормирование этих функций выполнялось с параметрами $T = 1$, $C = 15$, $\varepsilon = 0{,}1$.

Нелинейное по времени динамическое сопоставление данных, описанное в [6], было модифицировано для задачи сегментирования речи и выполняется по следующему алгоритму.

Пусть $\{\overline{S(n)}\} = \{S(0), S(1), \ldots, S(m), \ldots, S(N)\}$ есть последовательность векторов (спектральных срезов сонограммы) синтезированной речи, а $\{\overline{E(m)}\} = \{E(0), E(1), \ldots, E(m), \ldots, E(M)\}$ — последовательность векторов естественной речи.

Первым шагом является нахождение матрицы локальных расстояний $d\big(E(m),S(n)\big)$ между векторами синтезированной и естественной речи:

$$d\left(E(m),S(n)\right)=\sum_{i=1}^{C} Delta(E(m,i),S(n,i)),$$

$$Delta(E(m,i),S(n,i))=\begin{cases}\dfrac{\left|E(m,i)-S(n,i)\right|}{1-delta}, & \text{если} \quad \left|E(m,i)-S(n,i)\right|>delta,\\ 0, & \text{если} \quad \left|E(m,i)-S(n,i)\right|\le delta,\end{cases}$$

где $delta$ — эмпирический коэффициент, предназначенный для увеличения дикторонезависимости.

Следующим шагом является вычисление матрицы интегральных расстояний $D(m,n)$, матрицы времен $T(m,n)$ и матрицы переходов $Tr(m,n)$ со следующими начальными условиями:

$T(m,0)=0$, $T(0,n)=0$,

$D(m,0)=d\left(E(m),S(0)\right)$, $D(0,n)=d\left(E(0),S(n)\right)+D(0,n-1)+k_t*\left|n-1\right|$,

$Tr(m,0)=TrEnd$

для всех $m=\overline{1,M}$, $n=\overline{0,N}$.

Остальные значения $D(m,n)$, $T(m,n)$ и $Tr(m,n)$ рассчитываются в соответствии с рекуррентными формулами, приведенными ниже:

$$D(m,n)=\min\begin{bmatrix} D(m-1,n)+k_h d\left(E(m),S(n)\right)+Y\left(n,T(m-1,n)\right) & (1)\\ D(m,n-1)+k_v d\left(E(m),S(n)\right)+Y\left(n-1,T(m,n-1)\right) & (2)\\ D(m-1,n-1)+k_d d\left(E(m),S(n)\right)+Y\left(n-1,T(m-1,n-1)\right) & (3)\end{bmatrix}$$

$$T(m,n)=\begin{cases} T(m-1,n)+1, & \text{если} \quad D(m,n)=(1),\\ T(m,n-1), & \text{если} \quad D(m,n)=(2),\\ T(m-1,n-1)+1, & \text{если} \quad D(m,n)=(3).\end{cases}$$

$$Tr(m,n)=\begin{cases} TrHoriz, & \text{если} \quad D(m,n)=(1),\\ TrVert, & \text{если} \quad D(m,n)=(2),\\ TrDiag, & \text{если} \quad D(m,n)=(3).\end{cases}$$

$$Y(n,T)=\frac{k_t}{DestZero}\left|n-T\right|\cdot Path(n,T),\ Path(n,T)=\left(1-a^{-\left(\frac{n-T}{\frac{n+T}{2}}\right)^2}\right),\ a=0,5^{-\left(\frac{r-1}{\frac{r+1}{2}}\right)^{-2}},$$

где k_t — коэффициент учета времени; k_h, k_v, k_d — коэффициенты горизонтального, вертикального и диагонального перемещения соответственно; $Path(n,T)$ — функция, контролирующая возможность искажения оси времени синтезированного сигнала относительно оси времени размечаемого сигнала (в простейшем случае эта функция равняется 1, однако для задания возможного "коридора" искажения осей может быть использована указанная выше формула); r — коэффициент расширения "коридора" (на рис. 2 изображена функция $Path(n,T)$ с коэффициентом расширения коридора 0,75); $DestZero$ — нормирующий коэффициент, необходимый при поиске и сегментировании речи набором синтезированных эталонов, определяющий как минимум строки матрицы интегральных расстояний $D(m,N-1)$, при сопоставлении эталонной сонограммы с нулевой сонограммой (такой, в которой все значения равны 0); $TrHoriz$, $TrVert$, $TrDiag$ — некоторые различные числа, необходимые для последующего поиска обратного пути.

Значения коэффициентов k_t, k_h, k_v, k_d были подобраны экспериментальным методом и равнялись 0,01, 0,3, 0,9, 0,6, однако они могут и варьироваться.

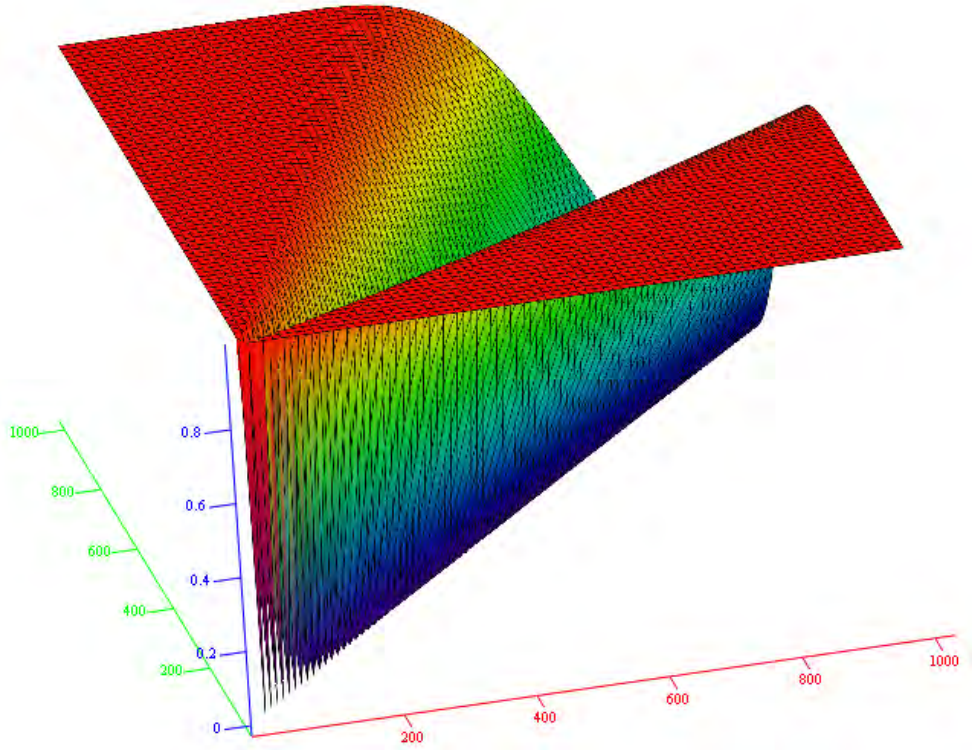

Рис. 2. Функция $Path(n,T)$ с коэффициентом расширения коридора 0,75

Сегментация данных заканчивается поиском минимума в последней строке матрицы интегральных расстояний $D(m,N-1)$, построении из этого минимума пути соответствия синтезированной и естественной сонограмм (путем анализа матрицы переходов $Tr(m,n)$) и переносом меток с синтезированного сигнала на естественный сигнал через найденный путь.

Автоматическое сегментирование речевого сигнала по сравнению с ручным отличается существенным увеличением скорости сегментации даже при контроле результатов работы оператором. Это преимущество может позволить в будущем создавать базы данных большего объема, для более качественного синтеза речи.

Однако описываемый в данной статье метод имеет и недостатки, связанные с тем, что граница между сегментами определяется исходя из сонограммы сигнала и может иметь погрешность ±10мс, а также с тем, что найденная граница не выровнена на границу питча. Дальнейшее развитие этого метода предполагается вести в направлении использования дополни-

тельных методов уточнения границ сегментов (например, метода линейного предсказания), а также использования методов расстановки питчей на вокализованных участках речи.

SPEECH SEGMENTATION ON THE BASE OF MODIFIED CONTINUOUS DYNAMIC PROGRAMMING

A.G. DAVYDAU, B.M. LOBANOV

Abstract

Construction of the system of segmentation of a speech signal is considered on the basis of a method of the modified continuous dynamic programming, with use the system of concatenation text to speech synthesis (TTS).

Литература

1. *Киселев В.В., Лобанов Б.М., Левковская Т.В., Хейдоров И.Э.* Тр. междунар. конф., посвященной 100-летию российской экспериментальной фонетики. СПб, 2001, С. 101–104.
2. *Давыдов А.Г., Киселев В.В., Лобанов Б.М., Цирульник Л.И.* // Изв. Белорус. инж. акад. 2004. № 1/1 С. 112–115.
3. *Hermansky H.* // J. Acoust. Soc. Am. 1990. Vol. 87, No. 4. P. 1738–1752.
4. *Hermansky H., Morgan N., Bayya A., Kohn P.* // Proc. EUROSPEECH. Genova, Italy. Sep. 1991. Vol. 3, P. 1367–1370.
5. *Logan B.* // Proc. International Symposium on Music Information Retrieval, Plymouth, MA, October. 2000.
6. *Вентцель Е.С.* Исследование операций: задачи, принципы, методология. М., 1988.