ИНФОРМАТИКА

2006 **январь-март** **№ 1**

УДК 004.934.2

А.Г. Давыдов

АЛГОРИТМ СЕГМЕНТАЦИИ РЕЧИ НА ОСНОВЕ МЕТОДА ДИНАМИЧЕСКОГО ПРОГРАММИРОВАНИЯ

Рассматривается система автоматической сегментации речи на основе динамического программирования. В качестве вектора признаков предлагается использовать спектр и усредненные конечные разности спектра по времени. Определяются оптимальные параметры работы системы на тестовом множестве из 1128 элементов.

Введение

Проблема сегментации речи в настоящее время является весьма актуальной в связи с широким спектром исследований по автоматическому синтезу речи по тексту и верификации речи диктора. Особенно остро данная проблема стоит для систем компиляционного синтеза речи по тексту, в которых количество и размер элементов синтеза напрямую влияют на качество создаваемой речи [1]. В связи с тем, что не в полной мере проведены экспериментальные исследования блока выделения признаков для задачи сегментации, целью работы является анализ блока выделения признаков и процедуры сопоставления, а также оценивание их оптимальных параметров. Решение данной задачи позволит сократить время, необходимое для разметки, и предоставить возможность включения блока сегментации в автоматические системы различного назначения [2].

1. Структура системы сегментации

В настоящее время для сегментации речи все более часто используется сопоставление методом динамического программирования обрабатываемого речевого сигнала с подобным ему синтезированным [3–5].

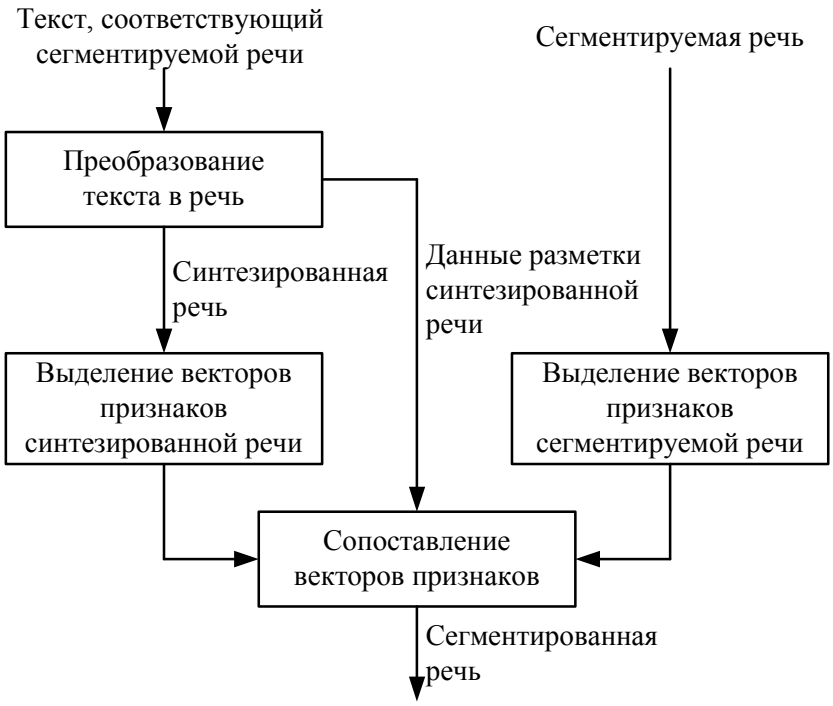

Рис. 1. Структура системы сегментации речи на основе метода динамического программирования

Система сегментации речи (рис. 1) включает в себя блок, реализующий преобразование текста в речи, необходимый для получения точных границ элементов синтеза в синтезированной речи. Данный блок имеет сложную структуру, и его функционирование широко исследовано в ряде работ [6–8].

2. Блок выделения признаков

Блок выделения признаков (рис. 2) предназначен для получения из речи наиболее информативного и устойчивого к искажениям в канале записи и изменению голоса диктора описания речевого сигнала. В качестве такого описания предлагается использовать спектр и усредненные конечные разности спектра (КРС) по времени. Это позволит выявить динамические свойства сигналов и повысить информативность признаков [9].

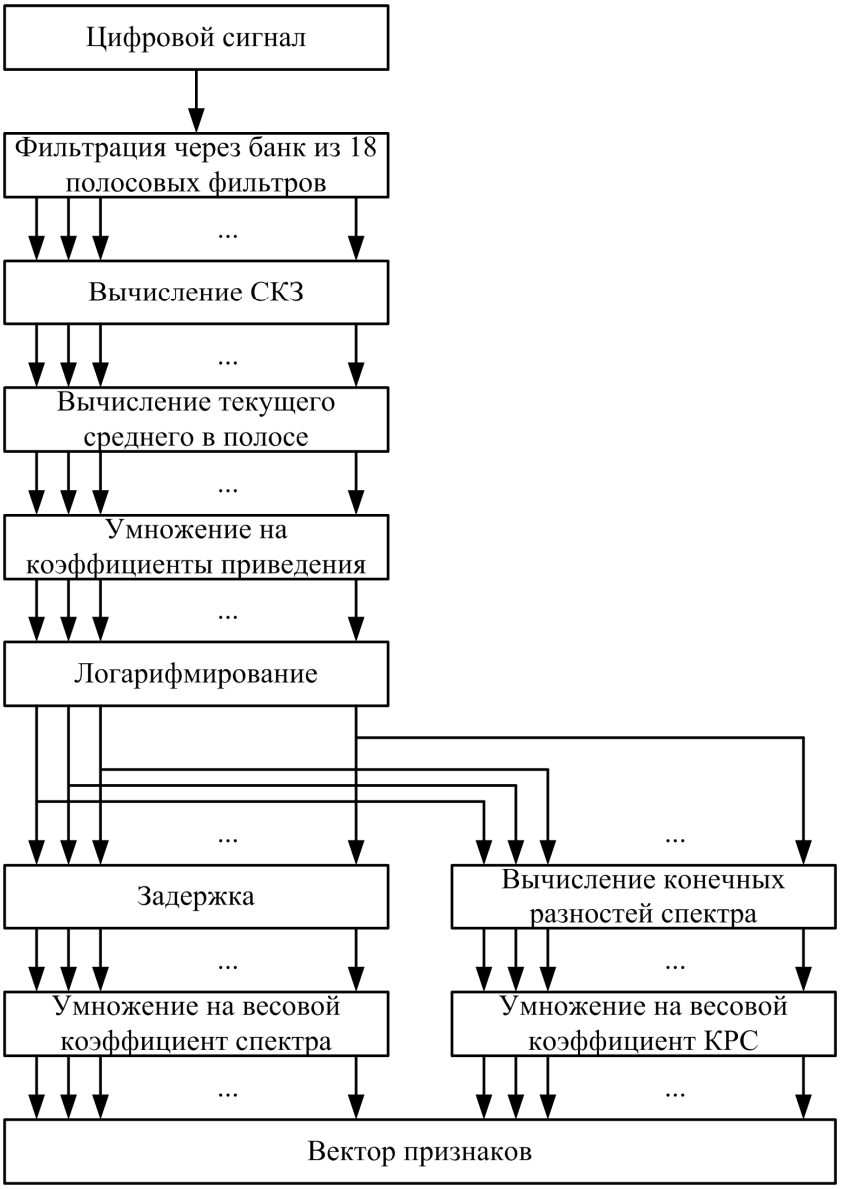

Рис. 2. Структурная схема блока вычисления вектора признаков

Для получения спектральной оценки сигнал пропускается через банк полосовых фильтров Баттерворта, являющихся аппроксимацией частотной шкалы барков [10], с перекрытием соседних полос в 2 барка для обеспечения лучшей дикторонезависимости (табл. 1).

Таблица 1

Полосы пропускания банка фильтров

Номер фильтра	1	2	3	4	5	6	7	8	9
Полоса пропускания, Гц	100 400	200 500	300 600	400 708	500 814	600 1000	708 1190	814 1415	1000 1682
Номер фильтра	10	11	12	13	14	15	16	17	18
Полоса пропускания, Гц	1190 2000	1415 2379	1682 2829	2000 3364	2379 4000	2829 4757	3364 5657	4000 6727	4757 8000

В каждом из 18 каналов вычисляется средний квадрат значения (СКЗ) сигнала на интервале дискретизации сонограммы по времени. Из полученной оценки амплитудного спектра находится текущее среднее при заданном интервале усреднения. С целью выделения из спектра наиболее значимых для сегментации полос, а также компенсации искажений, вносимых при использовании различных каналов записи, применяется умножение на коэффициенты приведения. Для приведения отличающихся по амплитуде сигналов к одному диапазону используется логарифмирование по основанию 10 [11, 12]. Вычисление конечных разностей спектра выполняется в каждом канале по следующей формуле:

$$\Delta S(i) = \frac{1}{K}\sum_{k=0}^{K-1}\left[S(i+k) - S(i-1-k)\right],$$

где $\Delta S(i)$ – усредненная конечная разность спектра по времени; K – интервал усреднения конечных разностей спектра; $S(i)$ – i-й спектральный отсчет по времени.

Интервалы дискретизации сонограммы, усреднения спектра и КРС являются переменными величинами. Определение их рабочих значений будет рассмотрено в разд. 4.

3. Блок сопоставления

Метод динамического программирования (ДП) широко применяется на практике [13–18], однако выбор числа точек анализа, их расположение и коэффициенты перехода из исходных точек в целевую меняются в зависимости от задачи. Наиболее часто используемым вариантом является трехточечная симметричная конфигурация. При этом коэффициенты переходов по горизонтали и вертикали равны единице, а коэффициент перехода по диагонали – единице или двум. При любой конфигурации для устранения грубых ошибок и ускорения вычислений все множество траекторий минимального расстояния следует ограничивать [15, 16, 18].

С целью детализации используемого метода ДП рассмотрим формулы, по которым производятся вычисления [14].

Пусть $\overline{E(m)} = \{E(0), E(1), \ldots, E(m), \ldots, E(M)\}$ – последовательность векторов признаков в эталонном слове (синтезированном в рассматриваемой задаче), а $\overline{S(n)} = \{S(0), S(1), \ldots, S(n), \ldots, S(N)\}$ – последовательность векторов в текущем речевом потоке, т. е. запись естественной речи, разбиваемая на элементы синтеза.

Первым шагом в процедуре ДП является нахождение матрицы локальных расстояний $d\left[\overline{S(n)}, \overline{E(m)}\right]$ между векторами эталона и текущего речевого потока:

$$d\left[\overline{S(n)}, \overline{E(m)}\right] = \sum_{l=1}^{L}\left|S(n,l) - E(m,l)\right|, \qquad (1)$$

где L – размерность векторов эталонного и текущего речевых потоков (в проведенных экспериментах использовались 36 каналов – 18 спектра и 18 каналов КРС).

Далее вычисляются матрицы интегральных расстояний $D(n,m)$, времен $T(n,m)$ и переходов $Tr(n,m)$. Начальные условия для расчетов следующие:

$$T(n,0)=0; \qquad T(0,m)=0;$$

$$D(n,0)=d\left[\overline{S(n)},\overline{E(0)}\right]; \qquad D(0,m)=d\left[\overline{S(0)},\overline{E(m)}\right]+D(0,m-1)+k|m-1|;$$

$$Tr(n,0)=TrEnd$$

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

Новые значения $D(n,m)$, $T(n,m)$ и $Tr(n,m)$ вычисляются в соответствие с формулами

$$D_H=D(n-1,m)+k_H d\left[\overline{S(n)},\overline{E(m)}\right]+\frac{k}{M}|m-T(n-1,m)|; \tag{2}$$

$$D_V=D(n,m-1)+k_V d\left[\overline{S(n)},\overline{E(m)}\right]+\frac{k}{M}|m-1-T(n,m-1)|; \tag{3}$$

$$D_D=D(n-1,m-1)+k_D d\left[\overline{S(n)},\overline{E(m)}\right]+\frac{k}{M}|m-1-T(n-1,m-1)|; \tag{4}$$

$$D(n,m)=\min\left[D_H,D_V,D_D\right]; \tag{5}$$

$$T(n,m)=\begin{cases} T(n-1,m)+1, & \text{если } D(n,m)=D_H; \\ T(n,m-1), & \text{если } D(n,m)=D_V; \\ T(n-1,m-1)+1, & \text{если } D(n,m)=D_D; \end{cases}$$

$$Tr(n,m)=\begin{cases} TrHoriz, & \text{если } D(n,m)=D_H; \\ TrVert, & \text{если } D(n,m)=D_V; \\ TrDiag, & \text{если } D(n,m)=D_D, \end{cases}$$

где k – весовой коэффициент времени; k_H, k_V, k_D – коэффициенты перехода по горизонтали, вертикали и диагонали соответственно.

Весовой коэффициент времени определяет степень возможного искажения временных шкал при поиске траектории минимального расстояния, при этом коэффициенты перехода по горизонтали, вертикали и диагонали определяют приоритетное направление – чем меньше значение соответствующего коэффициента, тем оно наиболее приоритетно. Матрица переходов содержит значения *TrHoriz*, *TrVert*, *TrDiag*, *TrEnd* и служит для нахождения временно□го соответствия между эталонным сигналом и анализируемым.

При ДП область поиска траектории соответствия между сравниваемыми сигналами ограничивается допустимым интервалом [15, 16, 18]. Ширина начала и ширина конца допустимого интервала являются переменными величинами, задаваемыми в виде доли от длины эталонного сигнала. Это позволяет увеличивать размер допустимого интервала для длинных фраз, где могут происходить значительные искажения временны□х шкал, и уменьшать его для коротких.

4. Оптимизация параметров системы сегментации

Для оценки точности определения границ фонем использовалась база из 94 записей, синтезированных голосами четырех дикторов (двух мужчин и двух женщин):

а	грибы	искры	легко	почки	Тася	утка
автор	губы	капать	марабу	пушка	тесто	Уфа
Ася	гуси	капот	молчи	пытка	тетка	цель
атом	густо	каратэ	мэтр	салют	тетя	чипсы
бабка	дата	каска	начеку	сев	тихий	шлюпка
батя	депо	Катя	не те	сети	топка	шутя
бублик	дочка	кафе	о	сито	тубус	э
будка	дудка	кладу	ода	сотка	Тэд	эпос
буква кэ	дядя	косит	опус	соха	тюлька	эти
быть	еда	кот	осень	Степа	тяга	этот
бэби	запад	кофта	пальто	судит	тяпка	
во-всю	зычный	кроты	пасха	сыпать	у	
выбор	и	кэб	плечо	сытый	убыл	
ВЭФ	Изя	Кэтрин	поза	тандем	усик	

Тестовое множество формировалось из пар всех возможных сочетаний произношения одинаковых фраз различными дикторами и составило 1128 элементов. Данные в базу записывались при частоте дискретизации 22 050 Гц и квантовании 16 бит/отсчет.

Процедура оптимизации параметров системы сегментации проводилась для целевой функции, состоящей из суммы среднего арифметического и СКЗ модуля ошибки сегментации.

На начальном этапе проведения исследований определяются исходные значения параметров (табл. 2).

Таблица 2

Начальные параметры тестирования системы сегментации

Параметры вычисления вектора признаков		Параметры динамического программирования	
Интервал дискретизации сонограммы	1 мс	Коэффициент горизонтального перехода	1
Интервал усреднения спектра	2 мс	Коэффициент вертикального перехода	1
Коэффициенты приведения	–40 дБ	Коэффициент диагонального перехода	1
Интервал усреднения КРС	14 мс	Весовой коэффициент времени	0
Весовой коэффициент спектра	1	Ширина начала допустимого интервала	0,03
Весовой коэффициент КРС	0	Ширина конца допустимого интервала	0,3

Для оценки необходимой величины *интервала усреднения спектра* анализировались его значения в диапазоне от 1 до 40 мс включительно с шагом 1 мс. Зависимость целевой функции от интервала усреднения спектра показана на рис. 3, а минимальное значение целевой функции оценено при отсутствии усреднения.

Важным параметром анализа является значение *коэффициентов приведения*, переносящее логарифмированные усредненные оценки амплитудного спектра из одного диапазона в другой. Для определения подходящей величины данных коэффициентов анализировался ряд их значений от –80 до 0 дБ с шагом в 1 дБ при интервале усреднения спектра, равном 1 мс, и значениями остальных параметров, указанными в табл. 2. Минимальное значение целевой функции наблюдалось при величине коэффициентов приведения, равной –48 дБ. Данная величина должна корректироваться при переходе к другому числу уровней квантования сигнала.

Величина *интервала усреднения КРС* тесно связана с интервалом усреднения спектра и весовым коэффициентом КРС. В связи с этим целесообразно провести анализ зависимости целевой функции от переменных интервала усреднения спектра и интервала усреднения КРС для нескольких значений коэффициента КРС. Для анализа использовался диапазон значений интервала усреднения спектра от 1 до 7 мс с шагом 1 мс, интервала усреднения КРС от 1 до 40 мс так же с шагом 1 мс и коэффициента КРС от 0,5 до 2 с шагом 0,5. Значения остальных параметров равнялись величинам, указанным в табл. 3, значение коэффициентов приведения – –48 дБ.

Минимальное значение целевой функции наблюдалось при коэффициенте КРС, равном 2, интервале усреднения спектра 1 мс (т. е. при отсутствии усреднения на интервале дискретизации сонограммы по времени) и интервале усреднения КРС, равном 15 мс (рис. 4).

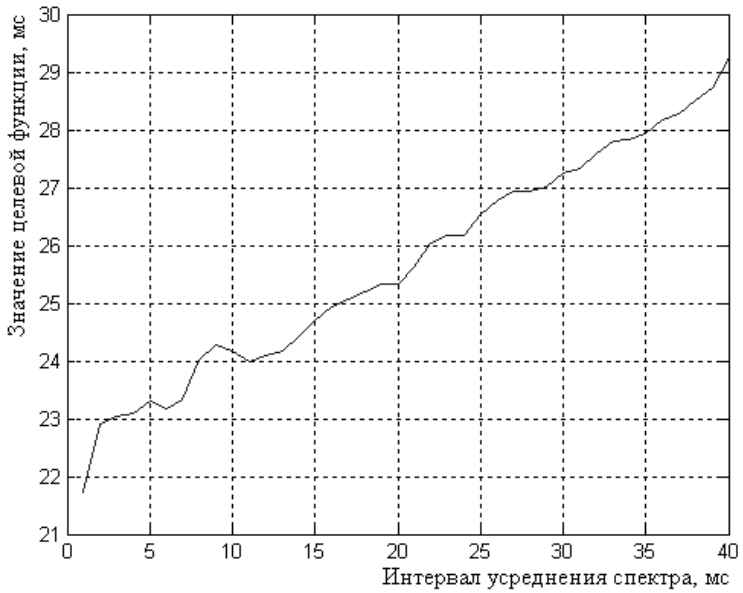

Рис. 3. Результаты анализа интервала усреднения

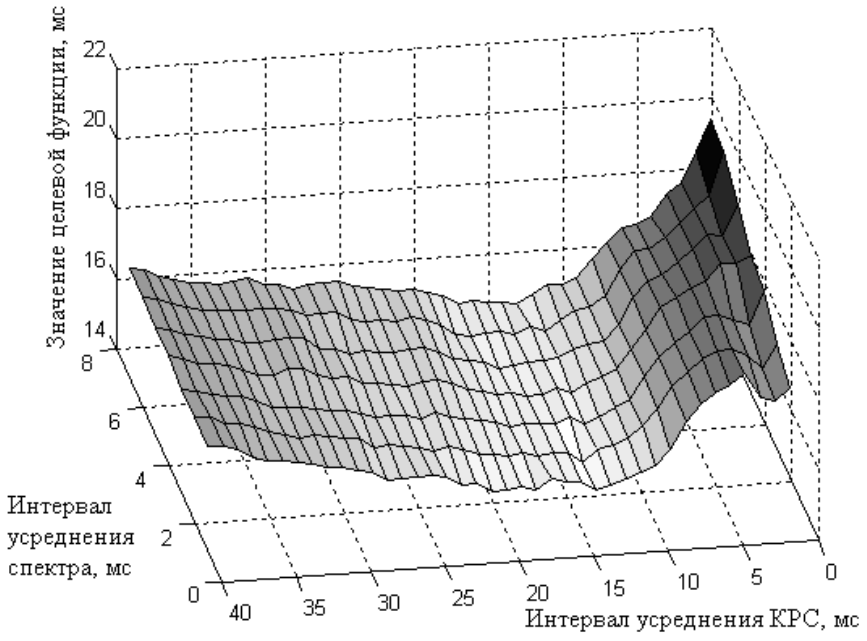

Рис. 4. Результаты анализа интервала усреднения КРС

Из формул (1)–(5) видно, что при выполнении процедуры ДП важным является не абсолютное значение весовых коэффициентов спектра и КРС, а их отношение. При поиске оптимальной величины *весового коэффициента КРС* анализировался диапазон значений интервала усреднения КРС от 12 до 16 мс с шагом 1 мс и весового коэффициента КРС от 5 до 15 с шагом 0,1 при весовом коэффициенте спектра, равном 1, интервале усреднения спектра 1 мс, коэффициентах приведения –48дБ и значениями остальных параметров, указанными в табл. 3 (рис. 5).

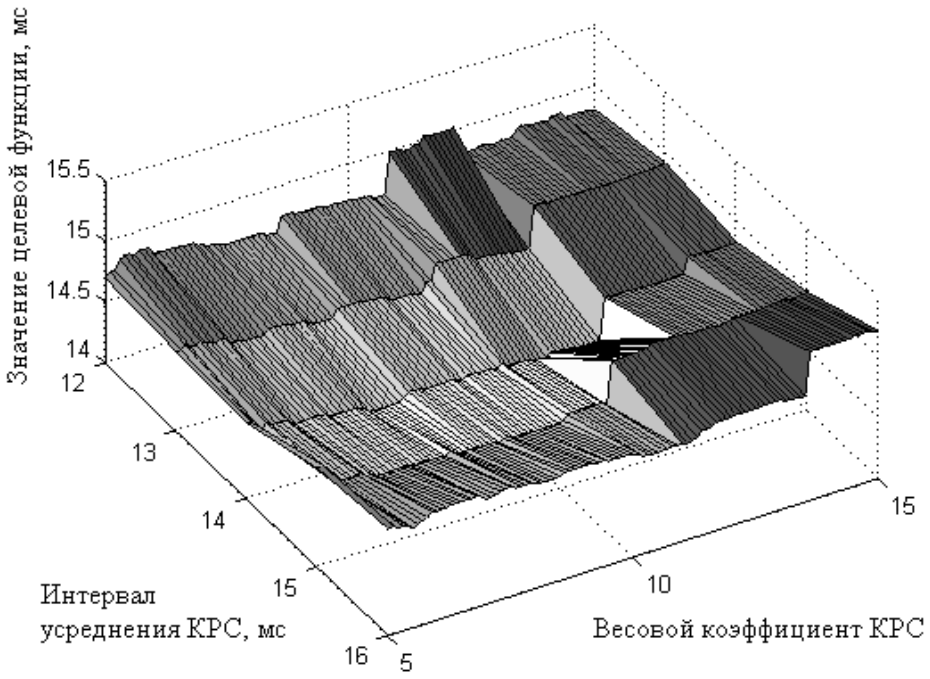

Рис. 5. Результаты анализа весового коэффициента КРС

Из проведенных исследований видно, что для дальнейшей оптимизации параметров работы системы сегментации целесообразно использовать ранее оцененные оптимальные параметры вычисления вектора признаков.

Как и в случае с весовыми коэффициентами спектра и КРС, из формул (2)–(5) видно, что при определении *коэффициентов горизонтального, вертикального и диагонального переходов* важным является только их отношение друг к другу. Зависимости целевой функции от изменения коэффициентов диагонального и вертикального переходов для различных значений коэффициента горизонтального перехода показаны на рис. 6. Величины коэффициентов изменялись на интервале от 0,1 до 1,2 с шагом 0,1. Минимальное значение целевой функции наблюдается при соотношении коэффициентов $k_H : k_V : k_D = 1:1:1$.

Для определения подходящего значения *весового коэффициента времени* анализировался интервал его значений от 0 до 2 с шагом 0,01. Результаты данного анализа показывают, что при значении весового коэффициента времени, находящегося в районе 1, сумма среднего арифметического модуля ошибки сегментации и его среднего квадрата принимает минимальное значение (рис. 7). Однако необходимо учесть, что при значительном расхождении в длительности фраз следует уменьшать значение данного коэффициента вплоть до нуля во избежание грубых ошибок в сегментации.

Для определения влияния параметров ширины начала и ширины конца допустимого интервала на точность разметки исследовалась зависимость значения целевой функции от указанных параметров в диапазоне от 0 до 0,5 включительно с шагом 0,05 (рис. 8).

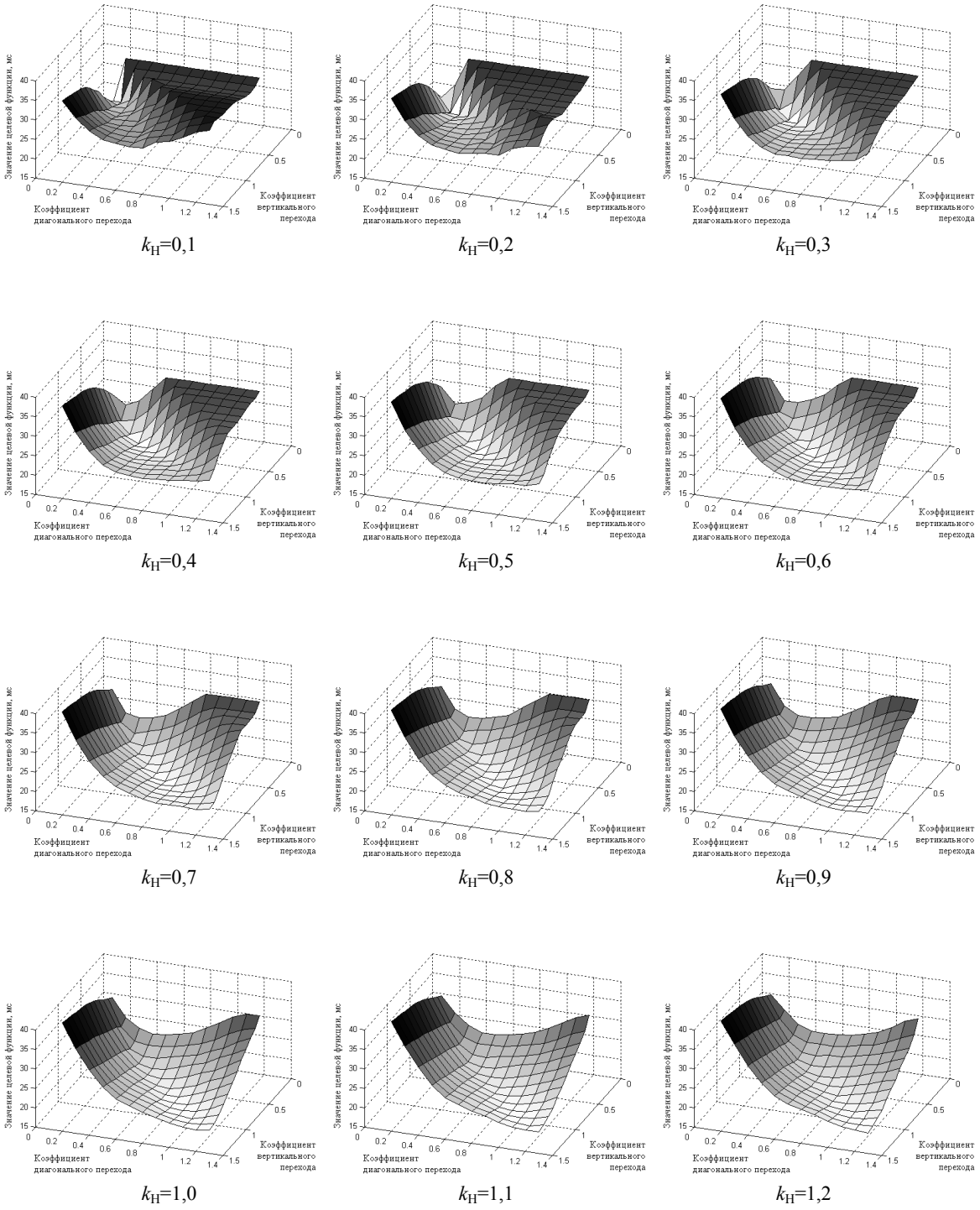

Рис. 6. Зависимость значения целевой функции при изменении коэффициента горизонтального перехода от коэффициентов вертикального и диагонального переходов

Минимальное значение целевой функции достигается при ширине начала допустимого интервала, равной 0, и ширине конца допустимого интервала, равной 0,2. Задание такого интервала позволяет избежать грубых ошибок в разметке. Однако оптимальные значения, вычисленные для некоторого тестового множества, могут послужить причиной искажения разметки при неравных временных задержках в начале сравниваемых сигналов. Во избежание ошибок предлагается выбирать ширину начала допустимого интервала, равную 0,05, и ширину конца, равную 0,30 – 0,35.

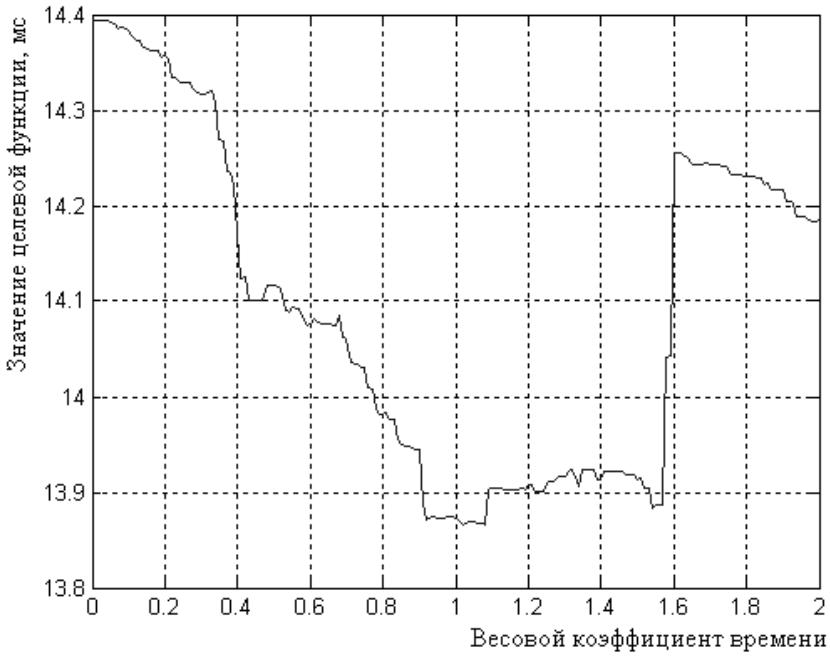

Рис. 7. Определение оптимального значения весового коэффициента времени

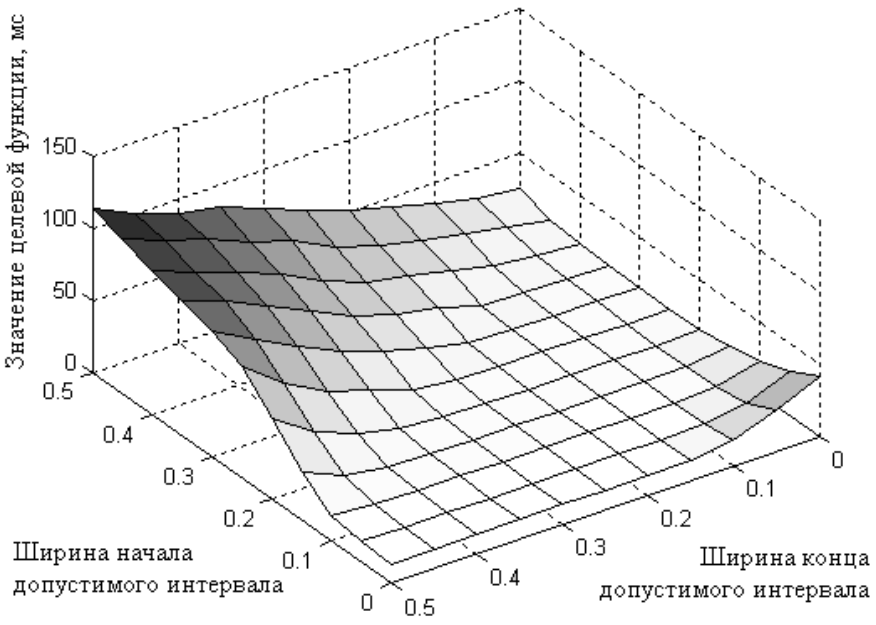

Рис. 8. Определение допустимого интервала при сопоставлении векторов признаков

В результате проведенного экспериментального анализа системы сегментации речи на основе метода динамического программирования, которая использует вектор признаков, состоящий из спектра и усредненных конечных разностей спектра по времени, определены оптимальные параметры работы системы сегментации речи (табл. 3).

Таблица 3

Оптимальные параметры работы системы сегментации речевого сигнала

Параметры вычисления вектора признаков		Параметры динамического программирования	
Интервал дискретизации сонограммы	1 мс	Коэффициент горизонтального перехода	1
Интервал усреднения спектра	1 мс	Коэффициент вертикального перехода	1
Коэффициенты приведения	–48 дБ	Коэффициент диагонального перехода	1
Интервал усреднения КРС	14 мс	Весовой коэффициент времени	1
Весовой коэффициент спектра	1	Ширина начала допустимого интервала	0,05
Весовой коэффициент КРС	8,1	Ширина конца допустимого интервала	0,35

Заключение

Использование в системе сегментации вектора признаков, состоящего из спектра и усредненных КРС по времени, позволило на тестовом множестве определить оптимальные параметры системы сегментации речи, а также уменьшить среднюю арифметическую ошибку расположения границ сегментов до 5,4 мс и ее среднее квадратическое значение до 10 мс.

Данное исследование было проведено при поддержке европейского фонда INTAS в рамках проекта «Разработка многоголосовой и многоязыковой системы синтеза и распознавания речи (языки: белорусский, польский, русский)» в соответствии с грантом INTAS № 04-77-7404.

Список литературы

1. Lobanov B.M, Karnevskaya E.B. Phonetics and its Applications. – Stuttgart: Franz Steiner Verlag, 2002. – P. 445–452.

2. Система экспресс-идентификации голоса личности методом клонирования акустических характеристик речи /А.Г. Давыдов, В.В. Киселев, Б.М. Лобанов, Л.И. Цирульник // Тез. докл. Междунар. конф. «Теория и практика речевой коммуникации». – М., 2004. – С. 23–28.

3. Malfrere F., Dutoit T. High quality speech synthesis for phonetic speech segmentation // Proc. of Eurospeech'97. – Rhodes, Greece, 1997. – P. 2631–2634.

4. Система сегментации речевого сигнала методом анализа через синтез / А.Г. Давыдов, В.В. Киселев, Б.М. Лобанов, Л.И. Цирульник // Известия Белорусской инженерной академии. – № 1 (17)/1'. – 2004. – С. 112–115.

5. Sethy A., Narayanan S. Refined speech segmentation for concatenative speech synthesis // Proc. of ICSLP 2002 – INTERSPEECH 2002. – Denver, USA, 2002. – P. 149–152.

6. Лобанов Б.М. Синтез речи по тексту // Четвертая Междунар. летняя школа-семинар по искусственному интеллекту: сб. науч. тр. – Мн.: Изд-во БГУ, 2000. – С. 57–76.

7. Development of an emotional speech synthesizer in Spamish / J.M. Montero, J. Guiterrez-Arriola, J. Colas et al. // Proc. of Eurospeech'99. – Budapest, Hungary, 1999. – P. 2099–2102.

8. Aravoice: An Arabic Text-to-Speech system / Z. Zemirli, R.A. Obrecht, A. Henni, M. Sellami // Proc. of SPECOM'2003. – Moskow, Russia, 2003. – P. 170–177.

9. Сорокин В.Н., Цыплухин А.И. Сегментация и распознавание гласных // Информационные процессы. – 2004. – Т. 4. – № 2. – С. 202–220.

10. Zwicker E., Flottorp G., Stevens S.S. Critical bandwidth in loudness summation // J. Acoust. Soc. Am. – № 29. – 1957. – P. 548–557.

11. Hermansky H., Morgan N. RASTA processing of speech // IEEE Trans. on Speech and Audio Proc. – 1994. – Vol. 2. – № 4. – P. 578–589.

12. A Low-Power, Fixed-Point, Front-End Feature Extraction for a Distributed Speech Recognition System / B. Delaney, N. Jayant, M. Hans et al. // IEEE International Conference on Acoustic Speech and Signal Processing, May 2002. – Orlando, Florida, 2002.

13. Bellman R.E. Dynamic Programming // Princeton University Press. – Princeton, NJ, USA, 1957.

14. Лобанов Б.М., Слуцкер Г.С., Тизик А.П. Автоматическое распознавание звукосочетаний в текущем речевом сигнале // Тр. НИИР. – Вып. 4. – М., 1969. – С. 67–75.

15. Itakura F. Minimum Prediction Residual Principle Applied to Speech Recognition // IEEE Transactions on Acoustics, Speech and Signal Processing. – Vol. ASSP-23. – 1975. – P. 52–72.

16. Sakoe H., Chiba S. Dynamic programming algorithm optimization for spoken word recognition // IEEE Transactions on Acoustics, Speech and Signal Processing. – Vol. 26. – 1978. – P. 43–49.

17. Вентцель Е.С. Исследование операций: задачи, принципы, методология. – М.: Наука, 1988. – 208 с.

18. Salvador S., Chan P. FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space // KDD Workshop on Mining Temporal and Sequential Data, August 22, 2004. – Seattle, Washington, 2004.

Поступила 15.08.05

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: andrew@ssrlab.com

A.G. Davydau

DYNAMIC PROGRAMMING SPEECH SEGMENTATION ALGORITHM

A system of automatic speech segmentation is considered on the basis of dynamic programming. The spectrum and averaged spectrum final differences on time are offered as a feature vector. The optimum parameters of the system are determined on the test set including 1128 elements.