Борису Мефодьевичу,
вызвавшему к жизни данный
метод, в знак дальнейшего, еще
более плодотворного сотрудничества.
09.IX.68.

bogdanov@vniisi.msk.su

УДК 621.391.192.5

Г.С.СЛУЦКЕР

НЕЛИНЕЙНЫЙ МЕТОД АНАЛИЗА РЕЧЕВЫХ СИГНАЛОВ*/

Работа посвящена поиску оптимальной метрики в пространстве речевых сигналов, учитывающей их временную структуру. Дается определение локальной и интегральной метрики.

При автоматическом распознавании речевых образов и фонетическом кодировании речи неизбежно возникает проблема выбора метрики в пространстве речевых сигналов. На этапе обучения (выработки эталонов) происходит сравнение двух реализаций одного или разных классов. Выбор метрики для сравнения двух реализаций должен приводить к увеличению компактности каждого класса. Эксперименты, проведенные рядом исследователей [1] , показывают, что общие линейные методы увеличения компактности классов, примененные непосредственно к мгновенным спектрам речевых сигналов, задачи не решают.

В работе [2] показывается неправомерность непосредственного среднеквадратичного сравнения двух спектральных сечений и приводится пример интегрального преобразования спектра в "фонетическую функцию", инвариантную относительно медленных мешающих изменений частотных характеристик речевого тракта и акустической среды:

$$P(\omega,t)=\frac{1}{T}\int_0^{\infty} e^{-\frac{\tau}{T}} \ln\frac{|S(\omega,t)|}{|S(\omega,t-\tau)|}\,d\tau \quad . \qquad (1)$$

Опираясь на преобразование (1), будем считать речевой сигнал заданным в виде $P(\omega,t)$, где ω – частота, t – текущее время. Так как функция (1) допускает среднеквадратичное сравнение сечений вдоль оси ω [2] , то расстояние между сечениями двух сигналов $P_1(\omega,t)$ и $P_2(\omega,t)$ в моменты времени t_1 и t_2 соответственно может быть измерено по формуле:

$$r(t_1,t_2)=\sqrt{\int_{\Omega}\left[P_1(\omega,t_1)-P_2(\omega,t_2)\right]^2 d\omega} \quad , \qquad (2)$$

где Ω – полоса частот; $P_1(\omega,t_1)$ – сечение сигнала P_1 в момент времени t_1; $P_2(\omega,t_2)$ – сечение сигнала P_2 в момент времени t_2 .

В связи с поисками в дальнейшем максимума некоторого функционала и для доказательства его существования нормализуем расстояние (2) между сечениями фонетических функций по формуле

$$q(t_1,t_2)=e^{-\alpha r(t_1,t_2)} \quad , \qquad (3)$$

где $\alpha>0$ – некоторая постоянная. Полученная функция (3) принимает значения на отрезке $[0,1]$ и может быть интерпретирована как мера подобия сечений $P_1(\omega,t_1)$ и $P_2(\omega,t_2)$. Назовем функцию (3) локальной мерой подобия двух сигналов P_1 и P_2 .

Пусть теперь T_1 и T_2 – соответственно длительности некоторых сигналов (звукосочетаний или слов). Ставится задача на основе построенной локальной меры подобия (3) и с учетом порядка следования подобных элементов определить

---

*/ Работа доложена на 5-й конференции молодых специалистов НИИР 25 октября 1967 г.

степень сходства двух данных сигналов произвольной конечной длительности (будем называть ее интегральной мерой подобия сравниваемых сигналов).

Рассмотрим функцию $q(t_1, t_2)$ ф-лы (3). Она определена в прямоугольнике $[T_1 \times T_2]$ плоскости времен (t_1, t_2) и принимает значения на отрезке $[0; 1]$ оси подобия q . На рис.1 изображена соответствующая поверхность $q(t_1, t_2)$, расположенная в параллелепипеде $[T_1 \times T_2 \times 1]$. Любая пара моментов времени (t_1°, t_2°) дает точку B плоскости времен, в которой после сравнения выбранных таким образом сечений сигналов $P_1(\omega, t_1^\circ)$ и $P_2(\omega, t_2^\circ)$ по ф-лам (2) и (3) можно вычислить меру подобия $q^\circ(t_1^\circ, t_2^\circ)$ этих сечений, а значит и точку $C(t_1^\circ t_2^\circ, q^\circ)$ указанной поверхности.

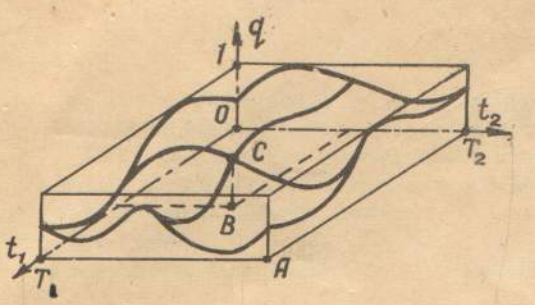

Рис. 1

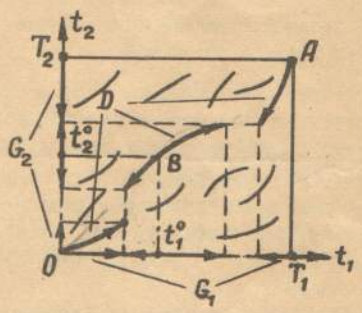

Рис.2

Итак, точка B плоскости времен ставит в соответствие друг другу моменты времени t_1 и t_2 двух процессов P_1 и P_2 , а функция $q(t_1, t_2)$ дает меру подобия соответствующих сечений. Проекции полюсов функции q образуют в плоскости времен отрезки кривых (рис.2), соединяя которые различным образом, мы получим множество $\{D\}$ траекторий соответствия (на рис.2 множества моментов времени, выбранные одной из траекторий D , обозначены через G_1 и G_2).

Задача построения интегральной меры подобия двух отрезков речи сведется к отысканию экстремума функционала, заданного на некотором классе траекторий. Значение функционала на данной траектории D должно зависеть не только от величины меры подобия $q(B)$ в каждой точке B траектории, но и от поведения этой кривой в окрестности точки B . Для учета локальных свойств кривой D в точке B введем некоторый неотрицательный вес $h_D(B)$. Искомый функционал принимает следующий вид:

$$F(D)=\int_D q(B)\, h_D(B)\, dB \,. \qquad (4)$$

С помощью (4) можно точнее сформулировать понятие интегральной меры подобия. А именно, интегральной мерой подобия назовем точную верхнюю грань

$$R=\sup F(D) \qquad (5)$$

значений функционала (4) на множестве допустимых траекторий соответствия $\{D\}$ двух данных отрезков речи.

Для определения класса допустимых траекторий $\{D\}$ и вида функции $h_D(B)$ разберем некоторые специфические свойства образования и восприятия речевых сигналов как временных процессов. Мы назовем эти свойства количественно-временными в отличие от качественных физических свойств речевых сигналов, учитываемых фонетической функцией (1), а также локальной мерой подобия $q(B)$, входящей под знак интеграла (4).

1. Свойство необратимости речевых процессов состоит в невозможности ставить в соответствие одновременно начало одного отрезка речи концу другого и наоборот – конец первого началу второго. Это значит, что две точки $B'(t_1', t_2')$ и $B''(t_1'', t_2'')$ плоскости времен могут принадлежать какой-либо одной траектории D только в том случае, когда одинаков порядок следования их проекций на оси времени:

$$t_1' < t_1'' \ , \ t_2' < t_2'' \ . \qquad (6)$$

Это свойство сильно ограничивает класс допустимых траекторий, а именно, допускаются только монотонно возрастающие траектории, отображающие взаимно-однозначно друг на друга подмножества G_1 и G_2 отрезков времени T_1 и T_2 двух сигналов (см.рис.2). Таким образом, выбор точки $B^o(t_1^o, t_2^o)$ в качестве элемента траектории D исключает возможность использования в построении этой траектории таких точек $B'(t_1', t_2')$, что $t_1' \leq t_1^o$, $t_2' \geq t_2^o$ или таких точек $B''(t_1'', t_2'')$, что $t_1'' \geq t_1^o$, $t_2'' \leq t_2^o$, где равенства не наступают одновременно (рис.3).

2. Свойство неравномерности течения речевых процессов состоит в непостоянстве скорости образования и длительности звучания отдельных звукосочетаний (вплоть до их полной редукции) при переходе от одной реализации к другой, в зависимости от контекста, темпа речи, диалектной принадлежности диктора и свойственного ему распределения длительностей звуков и переходных участков. Отсюда очевидна неправомочность априорного линейного сравнения двух отрезков речи, исходящего только из общих длительностей T_1 и T_2 сравниваемых сигналов и не учитывающего их локальных физических свойств. На рис.4 приведен пример линейного сопоставления (прямая OA) двух реализаций слова "вар", равносильного линейной нормализации по длительности. В результате "отображаются" друг в друга отмеченные качественно различные участки звуков "а" и "в", а" и "р". Пунктиром обозначена искомая оптимальная кривая.

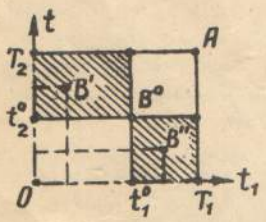

Рис.3

Свойство 2 окончательно определяет класс допустимых траекторий соответствия: допускаются нелинейные монотонно возрастающие кривые, могущие иметь разрывы в случае полной редукции (выпадения) отдельных звуков в одном из двух сигналов (см.рис.2). Ясно, что искомая интегральная мера подобия будет тем больше, чем больше мы сможем "нанизать" на одну монотонную кривую точек B, в которых $q(B)$ близко к 1, т.е. пар моментов времени, в которых два процесса наиболее сходны (с точки зрения выбранной локальной меры подобия q).

Осталось найти способ подсчета выбранных точек, т.е.весовую функцию h, которая бы учитывала взаимное расположение точек траектории и совместно с функцией совпадения качества q позволяла вычислить значение функционала F на данной траектории.

3. Свойство информативности количественно-временных характеристик речи приведет нас к конкретному выражению весовой функции $h_D(B)$. В самом деле, при ухудшении качества звучания речи большую информационную нагрузку берет на себя ритмическая структура сигнала (например, последовательность чередования ударных и безударных). С другой стороны, необходимо различать отрезки речи, отличающиеся ритмикой при одинаковом качественном составе

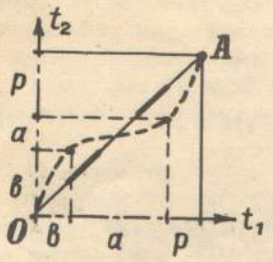

Рис.4

("замок" и "замок"). Значит, при сравнении двух сигналов необходимо требовать одновременно совпадения их качественного состава и ритмической структуры. Совпадение последней измеряется величиной нелинейного искажения траектории соответствия, т.е. весом $h_D(B)$.

Рассмотрим на рис.5 элемент dB траектории D , примыкающий к точке B . Если элемент dB параллелен биссектрисе 1-го координатного угла плоскости времен (назовем это направление главным), то его проекции dt_1 и dt_2 равны и естественной мерой "количества" точек, лежащих на элементе dB , будет их общая длина

$$h_D(B)dB = dt_1 = dt_2 . \qquad (7)$$

Если направление элемента dB отклоняется от главного (в любую сторону), то величина отклонения показывает разность скоростей течения двух процессов в сравниваемые моменты времени t_1 и t_2 . Отсюда вытекают следующие требования к весовой функции $h_D(B)$:

а) функция веса должна уменьшаться при увеличении отклонения направления элемента dB от главного;

б) знак отклонения не должен влиять на величину весовой функции, что равносильно условию равноправия двух процессов;

в) весовая функция должна непрерывно зависеть от величины отклонения.

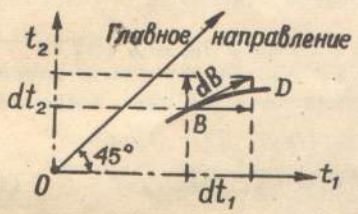

Рис.5

Перечисленные требования не определяют однозначно вид весовой функции, однако, если трактовать вопрос о ее отыскании как задачу измерения количества $h_D(B)dB$ точек, лежащих на элементе dB , то наиболее простой мерой будет выбор меньшей из двух проекций элемента dB (на рис.5 - dt_2):

$$h_D(B)dB = \min(dt_1 , dt_2). \qquad (8)$$

Формула (8) означает, что, если элемент dB ставит в соответствие друг другу отрезки времени dt_1 и dt_2 разной величины, то информация, общая для соответствующих отрезков сигнала, уже заключена в меньшем из них по длительности.

При равенстве проекций ф-ла (8) переходит в (7). Легко проверить выполнение требований а, б, в определенной по ф-ле (8) весовой функцией.

Кроме того, она обладает еще одним асимптотическим свойством:

г) вес $h_D(B)$ стремится к нулю при максимальном отклонении направления элемента dB от главного, т.е. при его стремлении к горизонтальному или вертикальному положению, когда одному моменту ставится в соответствие целый отрезок:

$$\lim_{\frac{dt_2}{dt_1} \to 0} h_D(B) = \lim_{\frac{dt_2}{dt_1} \to \infty} h_D(B) = 0 .$$

Подставив весовую функцию (8) в функционал (4), получим интеграл

$$F(D) = \int_D q(t_1, t_2) \min(dt_1, dt_2) ,$$

1. Свойство необратимости речевых процессов состоит в невозможности ставить в соответствие одновременно начало одного отрезка речи концу другого и наоборот – конец первого началу второго. Это значит, что две точки $B'(t_1', t_2')$ и $B''(t_1'', t_2'')$ плоскости времен могут принадлежать какой-либо одной траектории D только в том случае, когда одинаков порядок следования их проекций на оси времени:

$$t_1' < t_1'' \ , \ t_2' < t_2'' \ . \qquad (6)$$

Это свойство сильно ограничивает класс допустимых траекторий, а именно, допускаются только монотонно возрастающие траектории, отображающие взаимно-однозначно друг на друга подмножества G_1 и G_2 отрезков времени T_1 и T_2 двух сигналов (см.рис.2). Таким образом, выбор точки $B^\circ(t_1^\circ, t_2^\circ)$ в качестве элемента траектории D исключает возможность использования в построении этой траектории таких точек $B'(t_1', t_2')$, что $t_1' \leq t_1^\circ$, $t_2' \geq t_2^\circ$ или таких точек $B''(t_1'', t_2'')$, что $t_1'' \geq t_1^\circ$, $t_2'' \leq t_2^\circ$, где равенства не наступают одновременно (рис.3).

2. Свойство неравномерности течения речевых процессов состоит в непостоянстве скорости образования и длительности звучания отдельных звукосочетаний (вплоть до их полной редукции) при переходе от одной реализации к другой, в зависимости от контекста, темпа речи, диалектной принадлежности диктора и свойственного ему распределения длительностей звуков и переходных участков. Отсюда очевидна неправомочность априорного линейного сравнения двух отрезков речи, исходящего только из общих длительностей T_1 и T_2 сравниваемых сигналов и не учитывающего их локальных физических свойств. На рис.4 приведен пример линейного сопоставления (прямая OA) двух реализаций слова "вар", равносильного линейной нормализации по длительности. В результате "отображаются" друг в друга отмеченные качественно различные участки звуков "а" и "в", а" и "р". Пунктиром обозначена искомая оптимальная кривая.

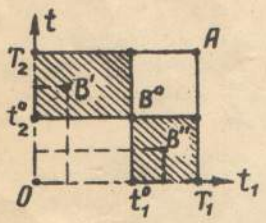

Рис.3

Свойство 2 окончательно определяет класс допустимых траекторий соответствия: допускаются нелинейные монотонно возрастающие кривые, могущие иметь разрывы в случае полной редукции (выпадения) отдельных звуков в одном из двух сигналов (см.рис.2). Ясно, что искомая интегральная мера подобия будет тем больше, чем больше мы сможем "нанизать" на одну монотонную кривую точек B, в которых $q(B)$ близко к 1, т.е. пар моментов времени, в которых два процесса наиболее сходны (с точки зрения выбранной локальной меры подобия q).

Осталось найти способ подсчета выбранных точек, т.е.весовую функцию h, которая бы учитывала взаимное расположение точек траектории и совместно с функцией совпадения качества q позволяла вычислить значение функционала F на данной траектории.

3. Свойство информативности количественно-временных характеристик речи приведет нас к конкретному выражению весовой функции $h_D(B)$. В самом деле, при ухудшении качества звучания речи большую информационную нагрузку берет на себя ритмическая структура сигнала (например, последовательность чередования ударных и безударных). С другой стороны, необходимо различать отрезки речи, отличающиеся ритмикой при одинаковом качественном составе

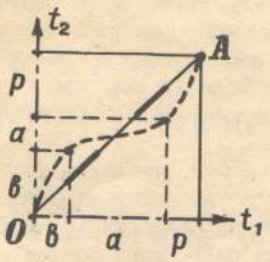

Рис.4

("замок" и "замок"). Значит, при сравнении двух сигналов необходимо требовать одновременно совпадения их качественного состава и ритмической структуры. Совпадение последней измеряется величиной нелинейного искажения траектории соответствия, т.е. весом $h_D(B)$.

Рассмотрим на рис.5 элемент dB траектории D , примыкающий к точке B . Если элемент dB параллелен биссектрисе 1-го координатного угла плоскости времен (назовем это направление главным), то его проекции dt_1 и dt_2 равны и естественной мерой "количества" точек, лежащих на элементе dB , будет их общая длина

$$h_D(B)dB = dt_1 = dt_2 . \qquad (7)$$

Если направление элемента dB отклоняется от главного (в любую сторону), то величина отклонения показывает разность скоростей течения двух процессов в сравниваемые моменты времени t_1 и t_2 . Отсюда вытекают следующие требования к весовой функции $h_D(B)$:

а) функция веса должна уменьшаться при увеличении отклонения направления элемента dB от главного;

б) знак отклонения не должен влиять на величину весовой функции, что равносильно условию равноправия двух процессов;

в) весовая функция должна непрерывно зависеть от величины отклонения.

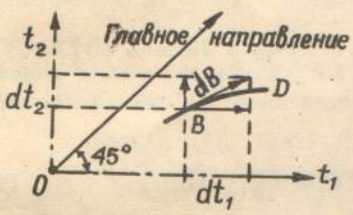

Рис.5

Перечисленные требования не определяют однозначно вид весовой функции, однако, если трактовать вопрос о ее отыскании как задачу измерения количества $h_D(B)dB$ точек, лежащих на элементе dB , то наиболее простой мерой будет выбор меньшей из двух проекций элемента dB (на рис.5 - dt_2):

$$h_D(B)dB = \min(dt_1 , dt_2). \qquad (8)$$

Формула (8) означает, что, если элемент dB ставит в соответствие друг другу отрезки времени dt_1 и dt_2 разной величины, то информация, общая для соответствующих отрезков сигнала, уже заключена в меньшем из них по длительности.

При равенстве проекций ф-ла (8) переходит в (7). Легко проверить выполнение требований а, б, в определенной по ф-ле (8) весовой функцией.

Кроме того, она обладает еще одним асимптотическим свойством:

г) вес $h_D(B)$ стремится к нулю при максимальном отклонении направления элемента dB от главного, т.е. при его стремлении к горизонтальному или вертикальному положению, когда одному моменту ставится в соответствие целый отрезок:

$$\lim_{\frac{dt_2}{dt_1} \to 0} h_D(B) = \lim_{\frac{dt_2}{dt_1} \to \infty} h_D(B) = 0 .$$

Подставив весовую функцию (8) в функционал (4), получим интеграл

$$F(D) = \int_D q(t_1, t_2) \min(dt_1, dt_2) ,$$

который сводится к общепринятым обозначениям

$$F(D)=\int_{D_1} q(t_1,t_2)\,dt_1+\int_{D_2} q(t_1,t_2)\,dt_2\,, \qquad (9)$$

где $$D_1=\left\{B\Big/\frac{dt_2}{dt_1}>1\right\}\,,\quad D_2=\left\{B\Big/\frac{dt_2}{dt_1}\le 1\right\}$$

два непересекающиеся подмножества точек B траектории D (см.рис.6).

Пользуясь свойством 1 [ф-ла (6)] монотонности траекторий и считая измеримыми исходные множества G_1 и G_2 (проекции кривой D), можно доказать [3] существование почти всюду на D касательных к ней, а значит, и возможность ее разбиения на две части D_1 и D_2. Кроме того, из монотонности функции на измеримом множестве следует измеримость ее производной [3], в частности, измеримость проекций E_1 и E_2 подмножеств D_1 и D_2 на оси t_1 и t_2 соответственно (рис.6). Будем считать далее подынтегральную функцию q измеримой на любой монотонной траектории. Тогда, вспомнив об ограниченности функции q

$$0\le q(t_1,t_2)\le 1\,, \qquad (10)$$

докажем существование конечных интегралов (9) в смысле Лебега [3].

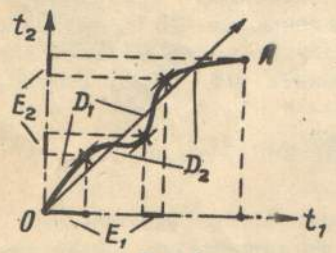

Рис.6

Из неравенств (10) вытекает также точная оценка значений функционала на данной траектории D:

$$0\le F(D)\le mE_1+mE_2\,, \qquad (11)$$

где mE_1 и mE_2 – меры проекций, соответствующих данной траектории D (см.рис.6). Правое равенство в (11) достигается только при $q\equiv 1$ на всей кривой D, за исключением, быть может, точек линейной меры нуль.

Учитывая свойства проекций E_1 и E_2, получим оценку значений функционала по всей совокупности допустимых траекторий:

$$0\le F\le \min(T_1,T_2)\,, \qquad (12)$$

где T_1 и T_2 – длительности сравниваемых сигналов. Из ограниченности (12) множества значений функционала (9) следует существование интегральной меры подобия (5) для любой пары сигналов.

В качестве метода вычисления меры (5) выберем метод "динамического программирования" [4]. Поскольку последний требует дискретизации по временным осям, то число допустимых траекторий будет конечно и хотя бы на одной из них функционал (9) достигнет максимального значения (5).

Разобьем сигналы P_1 и P_2 на временные интервалы равной длительности Δt (например, 0,01 сек). Проведем через точки деления прямые, параллельные осям t_1 и t_2 (рис.7). Усредним значения локальной меры подобия q в каждой полученной клетке и поставим в ней среднее значение q_{ij}, где i и j – номера интервалов Δt на осях t_1 и t_2, служащих проекциями этой клетки. Назовем

точки пересечения указанных прямых узлами и занумеруем их. Узел (i, j) будет точкой пересечения двух прямых $t_1 = i\Delta t$, $t_2 = j\Delta t$. Так как приближенная локальная мера подобия q_{ij} кусочно-постоянна в прямоугольнике $T_1 \times T_2$, то она терпит разрывы лишь в конечном числе точек на любой монотонной кривой, а значит, измерима на этой кривой. Перенумеруем узлы сети таким образом, чтобы координаты каждой клетки совпадали с координатами ее правого верхнего узла.

Можно показать, что оптимальная траектория будет состоять из отрезков прямых, соединяющих противоположные вершины одного элементарного квадрата. Учитывая асимптотическое свойство "г" весовой функции h, соединим между собой указанные диагонали квадратов, превратив траекторию в непрерывную ломаную линию. При этом интеграл (9) по отрезкам, параллельным осям координат, будем считать равным нулю (см.свойство "г"), а по диагоналям он будет равен $q_{ij}\Delta t$. Таким образом, в каждый узел $(i+1, j+1)$ мы можем придти тремя путями: 1) снизу вверх из узла $(i+1, j)$; 2) слева направо из узла $(i, j+1)$; 3) по диагонали клетки из узла (i, j). Только в третьем случае мы увеличим интеграл на величину $q_{i+1, j+1}\Delta t$.

Обозначим через F_{ij} значение функционала на оптимальном пути, соединяющем начало 0 с узлом (i, j). Тогда получим рекуррентную формулу

$$F_{i+1, j+1} = \max\left(F_{i+1, j};\ F_{i, j+1};\ F_{ij} + q_{i+1, j+1}\Delta t\right), \qquad (13)$$

позволяющую вычислить оптимальное значение функционала F в узле $(i+1, j+1)$, если найдены числа в узлах $(i+1, j)$, $(i, j+1)$ и (i, j). Будем пользоваться ф-лой (13) в следующем порядке. В узлах, расположенных на осях координат, положим $F_{i0} = F_{0j} = 0$. Затем вычисляем

$F_{11} = \max(F_{10}, F_{01}, F_{00} + q_{11}\Delta t)$ и т.д. для всех F_{1j} по 1-й вертикали вверх. Затем возвращаемся вниз на 2-ю вертикаль [узел (2,1)] и т.д. до точки A, в которой получим искомую интегральную меру подобия.

На рис.7 приведен числовой пример, иллюстрирующий описанный выше метод отыскания интегральной меры подобия двух сигналов. Полученное значение функционала (9) подчиняется оценкам (11) и (12). При этом правая часть (11) означает число диагоналей квадратов, вошедших в оптимальную траекторию

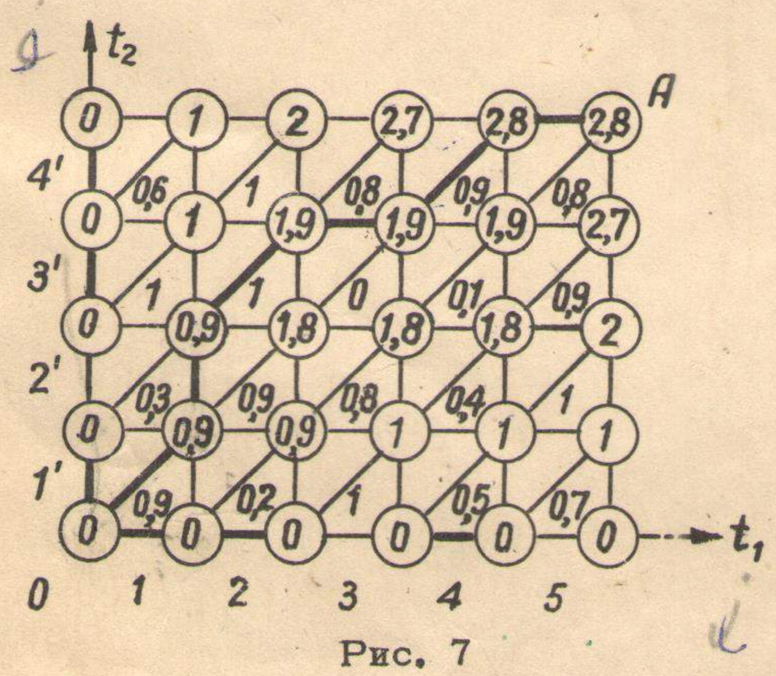

Рис. 7

$$mE_1 = mE_2 = 3 ,$$

в правой части (12) имеем

$$\min(T_1, T_2) = \min(5; 4) = 4 ;$$

Таким образом, получаем

$$2{,}8 < 3 < 4 .$$

Найденная на рис.7 оптимальная траектория автоматически ставит в соответствие друг другу пары 1′–1, 3′–2, 4′–4 отрезков сигналов и выбрасывает отрезки 3 и 5 из первого и отрезок 2′ из второго сигнала.

В заключение автор считает своим долгом выразить благодарность А.А.Пирогову, а также Л.В.Златоустовой и Б.М.Лобанову, участвовавшим в обсуждении ряда вопросов, разобранных в статье.

ЛИТЕРАТУРА

1. Б.Н.Епифанцев. О спектрах речевых сигналов. Труды Ленинградского Политехнического института им. М.И.Калинина. 1967, № 275.

2. А.А.Пирогов. К вопросу о фонетическом кодировании речи. "Электросвязь", 1967, № 5.

3. И.П.Натансон. Теория функций вещественной переменной. ГИТТЛ, 1950.

4. Е.С.Вентцель. Элементы динамического программирования. "Наука", 1964.

Статья поступила в редакцию 10 января 1968 г.