

УДК 004.934+004.8

Г. В. Дорохина

Государственное учреждение «Институт проблем искусственного интеллекта», г. Донецк
83048, г. Донецк, ул. Артёма, 118-б

МОДИФИКАЦИЯ АЛГОРИТМА DTW ДЛЯ ПОФОНЕМНОГО РАСПОЗНАВАНИЯ СЛОВ

G. V. Dorokhina

Public institution «Institute of Problems of Artificial intelligence», Donetsk
83048, c. Donetsk, 118 "b" Artjoma st.

MODIFICATION OF THE ALGORITHM DTW FOR SPOKEN WORD RECOGNITION BASED ON PHONEME RECOGNITION

Г. В. Дорохіна

Державна установа «Інститут проблем штучного інтелекту», м. Донецьк
83048, м. Донецьк, вул. Артема, 118-б

МОДИФІКАЦІЯ АЛГОРИТМУ DTW ДЛЯ ПОФОНЕМНОГО РОЗПІЗНАВАННЯ СЛІВ

Обоснована возможность пофонемной модификации алгоритма DTW. Предложен функционал для вычисления меры расхождения между фонетической транскрипцией слова и распознаваемым сигналом. Разработано представление словаря распознавания. Они составляют основу предложенного метода пофонемного распознавания. В проведенном эксперименте метод пофонемного превзошёл алгоритм DTW по частоте правильного распознавания и показал большую устойчивость к смене звукозаписывающего оборудования.

Ключевые слова: распознавание речи, алгоритм DTW, пофонемное распознавание.

Article demonstrate the possibility to modify algorithm DTW in order to recognize spoken words phoneme by phoneme. The functional for the calculation of the differences between the phonetic transcription of word and speech signal is proposed. The presentation of recognition vocabulary is developed. The proposed functional and developed presentation of recognition vocabulary form the basis of the proposed method for spoken word recognition speech recognition that uses phoneme recognition. This method surpassed algorithm DTW frequency of correct recognition and showed greater resistance to change recording equipment.

Key words: speech recognition, algorithm DTW, phoneme recognition.

Обґрунтовано можливість пофонемної модифікації алгоритму DTW. Запропоновано функціонал для обчислення міри розбіжності між фонетичною транскрипцією слова і сигналом, який розпізнають. Розроблено представлення словника розпізнавання. Вони складають основу запропонованого методу пофонемного розпізнавання. У проведеному експерименті метод пофонемного перевершив алгоритм DTW по частоті правильного розпізнавання і показав більшу стійкість до зміни звукозаписувального обладнання.

Ключові слова: розпізнавання мовлення, алгоритм DTW, пофонемне розпізнавання.

Важная роль в решении задач распознавания речи принадлежит методу Dynamic Time Warping (DTW), интерпретируемому как нелинейное растяжение-сжатие оси времени. Он являлся доминирующей парадигмой для распознавания изолированных слов с малым словарём распознавания. DTW в своей области применения давал очень хорошие результаты и фактически превосходил скрытые модели Маркова. От DTW по существу отказались из-за следующих проблем: введение модели языка не было естественным; задача построения синтетических эталонов осталась нерешенной; не найдено единой статистической формулировки распознавания, включающей все модули распознавания речи [1 с. 53-54].

Известно также о других проблемах метода распознавания по алгоритму DTW. Проблема большого разброса длин эталонов состоит в следующем. Если длина одного из эталонных сигналов значительно меньше длин остальных, мера расхождения от него до распознаваемого сигнала будет минимальной. Другая проблема метода формулируется так: «корректное временное выравнивание двух произнесений различных слов не является чётко определённой лингвистической концепцией» [2, с. 225].

Настоящая статья является развитием работы [3] и продолжает усилия коллектива под руководством В. Ю. Шелепова (ГУ «Институт проблем искусственного интеллекта») по решению проблемы синтетических эталонов – обучения небольшого набора базовых речевых единиц, более мелких, чем распознаваемые, и распознавания на их основе речевых единиц заданного словаря. Такой более мелкой единицей может выступать фонема – группа звуков, обладающих определенной артикуляторно-слуховой общностью и функционально друг другу не противопоставленных, т.е. не встречающихся в одинаковом положении в фонетической структуре слова [4].

В ГУ «Институт проблем искусственного интеллекта» успешно развивается метод дифонного распознавания [5-7], где в качестве базовых единиц используются дифоны. К настоящему моменту метод дифонного распознавания позволяет создавать системы распознавания слов со словарём в десятки тысяч слов. В то же время представляет интерес модификация алгоритма DTW, которая позволяла бы использовать образ фонемы целиком, включая и стационарную часть, и межфонемный переход.

Реализации фонемы, её варианты, обусловленные конкретным фонетическим окружением, называют аллофонами. Последовательность аллофонов слова является связующим звеном между звучанием и написанием слова. Символьное обозначение последовательности аллофонов слова называют фонетической транскрипцией.

Целью работы является развитие методов распознавания речи за счёт разработки метода пофонемного распознавания слов на основе алгоритма DTW.

Для достижения цели сформулированы и решены **следующие задачи**: обоснование возможности пофонемной модификации алгоритма DTW; разработка метода пофонемного распознавания; численное исследование программно-реализованного метода.

Обоснование возможности пофонемной модификации алгоритма DTW

Алгоритм DTW позволяет вычислить меру расхождения между эталоном E и распознаваемым сигналом R , представленными последовательностями векторов признаков:

$$E = e_1, e_2, \dots, e_i, \dots, e_n, \quad (1)$$

$$R = r_1, r_2, \dots, r_j, \dots, r_m. \quad (2)$$

Вычисление меры расхождения $d(E, R)$ производят путём вычисления матрицы расстояний D между векторами эталона e_i и распознаваемого сигнала r_j

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix}, \quad d_{ij} = d(e_i, r_j). \quad (3)$$

Далее, по матрице D вычисляют элементы DTW-матрицы K

$$K = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1m} \\ k_{21} & k_{22} & \cdots & k_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ k_{n1} & k_{n2} & \cdots & k_{nm} \end{bmatrix}, \quad (4)$$

по рекуррентным формулам:

$$\begin{aligned} k_{11} &= d_{11}, \\ k_{1j} &= d_{1j} + k_{1(j-1)}, \quad j \in [2; m], \\ k_{i1} &= d_{i1} + k_{(i-1)1}, \quad i \in [2; n], \\ k_{ij} &= d_{ij} + \min(k_{(i-1)(j-1)}, k_{(i-1)j}, k_{i(j-1)}), \quad i, j \in [2; n]. \end{aligned} \quad (5)$$

Меру расхождения между эталоном E и сигналом R определяют как:

$$d(E, R) = k_{nm}. \quad (6)$$

Кроме определения меры расхождения между эталоном E и сигналом R , полученная DTW-матрица K позволяет определить множество M пар номеров (p, q) соответствующих друг другу векторов e_i и r_j [4]:

$$M = \{(p_h, q_h)\}, \quad h \in \overline{[1, H]}, \quad (7)$$

где H – количество соответствующих друг другу пар векторов. Эти пары определяются итеративно по формулам:

$$\begin{aligned} p_H &= n, \quad q_H = m, \\ (p_{h-1}, q_{h-1}) &= \arg \min_{(p', q') \in P_h} (k_{p', q'}), \end{aligned} \quad (8)$$

где $P_h = \{(p', q') : p_{h-1} \leq p' \leq p_h, q_{h-1} \leq q' \leq q_h, (p' \neq p_h \vee q' \neq q_h)\}$

$$p_1 = 1, \quad q_1 = 1.$$

Множество M состоит из пар индексов (p_h, q_h) таких, что элементы $k_{p_h q_h}$ матрицы K принимают участие в формировании k_{nm} . Если обратиться к рекуррентным формулам (5), использованным для формирования значений матрицы K , видно, что мера расхождения $d(E, R)$ представима в виде суммы

$$d(E, R) = \sum_{h=1}^H d(e_{p_h}, r_{q_h}). \quad (9)$$

Покажем, что возможно модифицировать алгоритм DTW таким образом, что станет возможным фонемное сопоставление двух различных произнесений одного и того же слова без изменения меры расхождения между ними.

Пусть сигналы E и R – параметрические представления различных произнесений одного и того же слова. Известна фонетическая транскрипция $T = (t_1, \dots, t_i, \dots, t_l)$, обозначающая последовательность аллофонов этого слова. Обозначать аллофоны будем с помощью символов алфавита A . Дан вектор границ $B = (b_1, \dots, b_i, \dots, b_l)$ аллофонов в эталоне E , где b_i – номер первого вектора признаков e_{b_i} в сигнале E , который принадлежит i -му аллофону t_i . По вектору B и множеству M получим B' – вектор границ аллофонов в сигнале R :

$$B' = (b'_1, \dots, b'_i, \dots, b'_l), \quad (10)$$

где $\forall i = \overline{1, l} \quad b'_i = q_k : (p_k, q_k) \in M, p_k = b_i$.

Обозначим через $E[x; y]$ последовательность векторов признаков $e_x, e_{x+1}, \dots, e_y$, сигнала E . Будем называть $E[x; y]$ фрагментом сигнала E . Аналогично для $R[x'; y']$. Можем говорить о соответствии фрагментов $E[b_i; b_{i+1}-1]$ и $R[b'_i; b'_{i+1}-1]$, представляющих аллофон t_i в сигналах E и R . Это даёт возможность говорить о нахождении меры расхождения между E и R путём сопоставления их фрагментов, соответствующих отдельным аллофонам. Действительно, величина (9) может быть представлена как сумма мер расхождения между участками эталонного и распознаваемого сигнала, соответствующими одному и тому же аллофону:

$$d(E, R) = \sum_{i=1}^l \sum_{x=b_i}^{b_{i+1}-1} d(e_{p_x}, r_{q_x}), \quad (11)$$

считая, что $b_{l+1} = n+1$ и $b'_{l+1} = m+1$.

То есть меру расхождения $d(E, R)$ можно вычислять фонемно, если известен путь выравнивания M . Наличие информации о фрагменте пути выравнивания до конца i -го аллофона без вычисления всей DTW-матрицы позволило бы сократить вычислительные затраты, так как потребовалось бы вычислять не всю матрицу K , а только области, соответствующие границам каждого аллофона в эталонном и распознаваемом сигнале – подобно приведенному на рис. 1, где вычислению подлежат только незаштрихованные ячейки матрицы.

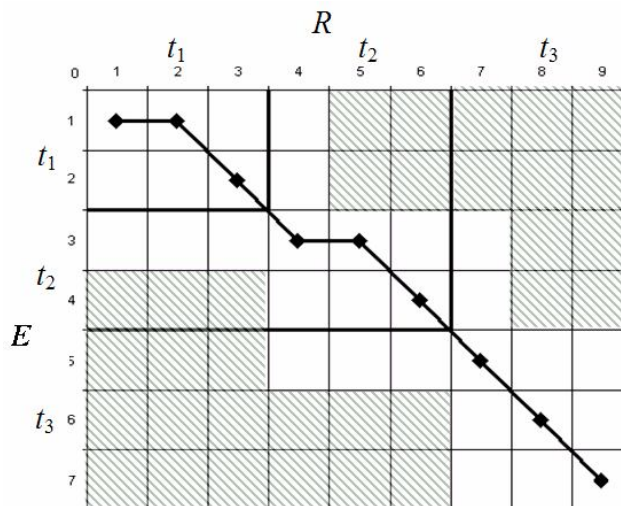


Рисунок 1 – Пример вычисления DTW-матрицы фонемно

Покажем, что это возможно, разработав алгоритм определения элементов множества M при прямом прохождении DTW-матрицы K .

Наполняя множество M , будем перемещаться от ячейки с индексами $(1, 1)$ к ячейке с индексами (n, m) .

Могут возникнуть ситуации, когда в процессе выполнения алгоритма потребуется перестроить фрагмент пути выравнивания путём замены нескольких последних элементов этого множества другими. Для правильной работы алгоритма необходимо потребовать, чтобы элементы во множестве M были перенумерованы таким образом, что пара $(p_i, q_i) \in M$ с большим порядковым номером i соответствовала элементу $k_{p_i q_i}$ с большим значением. То есть:

$$\forall i > j \quad k_{p_i q_i} > k_{p_j q_j},$$

где (p_i, q_i) и (p_j, q_j) – элементы пути выравнивания M , а $k_{p_i q_i}$ и $k_{p_j q_j}$ – элементы DTW-матрицы.

Алгоритм поиска пути выравнивания при прямом проходе DTW-матрицы начинаем с добавления в него пары индексов $(1, 1)$. Для краткости будем называть элемент DTW-матрицы, индексы которого хранятся в элементе множества M с наибольшим порядковым номером h , рабочим элементом, а саму эту пару индексов – индексами рабочего элемента.

Пусть $k_{p_h q_h}$ – рабочий элемент. Исходя из алгоритма вычисления элементов DTW-матрицы, элемент $k_{(p_h+1)(q_h+1)}$ равен DTW-расстоянию между фрагментами эталонного и распознаваемого сигналов $E[1, p_h + 1]$ и $R[1, q_h + 1]$, соответственно:

$$k_{(p_h+1)(q_h+1)} = d(E[1, p_h + 1], R[1, q_h + 1]).$$

В формировании значения $k_{(p_h+1)(q_h+1)}$ принимает участие либо рабочий элемент $k_{p_h q_h}$, либо один из соседних с ним элементов: $k_{(p_h+1)q_h}$ или $k_{p_h(q_h+1)}$. По отношению $k_{(p_h+1)(q_h+1)}$ элемент, формирующий его значение, принадлежит пути выравнивания фрагментов сигналов $E[1, p_h + 1]$ и $R[1, q_h + 1]$. Это следует из алгоритма поиска пути выравнивания при обратном проходе DTW-матрицы (8).

Проверив, совпадает ли рабочий элемент $k_{p_h q_h}$ с элементом, используемым при формировании значения $k_{(p_h+1)(q_h+1)}$, мы выясняем принадлежит ли рабочий элемент пути выравнивания фрагментов сигналов $E[1, p_h + 1]$ и $R[1, q_h + 1]$. Подтвердив принадлежность элемента $k_{p_h q_h}$ пути выравнивания, предполагаем принадлежность пути выравнивания элемента $k_{(p_h+1)(q_h+1)}$. Для этого добавляем в путь выравнивания M пару $(p_h + 1, q_h + 1)$, таким образом назначая $k_{(p_h+1)(q_h+1)}$ рабочим элементом.

Если же в формировании значения $k_{(p_h+1)(q_h+1)}$ принимает участие не рабочий элемент, то последний не может принадлежать пути выравнивания. Следовательно, путь выравнивания необходимо корректировать. Для этого формируем множество M' – корректирующий фрагмент пути выравнивания. Потребуем такой упорядоченности элементов множества M' , чтобы элементы с большим номером являлись индексами элемента DTW-матрицы с меньшим значением.

Строить корректирующий фрагмент пути выравнивания M' будем путём обратного прохода DTW-матрицы от элемента, сформировавшего значение $k_{(p_h+1)(q_h+1)}$, до тех пор, пока очередной найденный элемент (i_f, j_f) корректирующего фрагмента пути выравнивания M' не окажется принадлежащим пути выравнивания M . Допустим, g – наибольший номер элемента (p_g, q_g) пути выравнивания M , совпадающий с найденным элементом (i_f, j_f) .

Корректируем путь выравнивания M , удалив из него элементы с номерами $c = g + 1, h$ и затем добавив элементы корректирующего фрагмента пути выравнивания M' .

Описанные действия с рабочим элементом повторяем итеративно, пока индексами рабочего элемента не станет пара (n, m) .

Во всех экспериментах путь выравнивания M , определяемый по приведенному алгоритму (рис. 2) совпадает с получаемым при обратном проходе DTW-матрицы согласно (8).

1. Инициализация

$$p_1 := 1, q_1 := 1, M := \{(p_1, q_1)\}$$

2. Итеративное определение пары номеров соответствующих друг другу векторов эталонного и распознаваемого сигнала на $(h+1)$ -ом шаге:

2.1 Определяем пару индексов (p_{h+1}, q_{h+1}) , таких что

$$(\hat{p}, \hat{q}) := \arg \min_{(p', q') \in P_h} (k_{p' q'}),$$

где $P_h = \{(p', q') : p_h \leq p' \leq p_h + 1, q_h \leq q' \leq q_h + 1, p' \leq n, q' \leq m, p' = p_h \vee q' = q_h\}$;

2.2 Если $(\hat{p}, \hat{q}) = (p_h, q_h)$

2.2.1 то $(p_{h+1}, q_{h+1}) := (\hat{p} + 1, \hat{q} + 1)$ $M := M \cup (p_{h+1}, q_{h+1})$

2.2.2 иначе Пересчитать путь выравнивания M , от элемента с индексами $(\hat{p}, \hat{q})$, до первого элемента, принадлежащего пути выравнивания M

2.2.2.1 $f := 1, (i_f, j_f) := (\hat{p}, \hat{q}), M' := \emptyset$

2.2.2.2 Пока $(i_f, j_f) \notin M$ выполнять

2.2.2.2.1 $M' := M' \cup (i_f, j_f)$

2.2.2.2.2 $(i_{f+1}, j_{f+1}) := \arg \min_{(i', j') \in I_f} (k_{i' j'}),$

где $I_f = \{(i', j') : i_f - 1 \leq i' \leq i_f, j_f - 1 \leq j' \leq j_f, 1 \leq i', 1 \leq j', i' = i_f - 1 \vee j' = j_f - 1\}$

2.2.2.2.3 $f := f + 1$

2.2.2.3 Найдём g , такое что $(p_g, q_g) \in M, (p_g, q_g) = (i_f, j_f)$

2.2.2.4 $M := M \setminus (p_c, q_c), c = \overline{g+1, h}$

2.2.2.5 $M := M \cup M', h := |M|$

3. Условие останова: $p_{h+1} = n, q_{h+1} = m.$

4. Результат: $M = \{(p_h, q_h)\}_{h=1}^H$

Рисунок 2 – Поиск пути выравнивания при прямом проходе DTW-матрицы

Перейдём к разработке алгоритма (рис. 3) вычисления меры расхождения между образом непоследнего аллофона t_i слова с транскрипцией $T=(t_1, \dots, t_i, \dots, t_l)$ и соответствующего ему фрагмента распознаваемого сигнала $R[b_i, b_{i+1}-1]$.

Пусть i -й и $(i+1)$ -й аллофоны слова принимают значения g и h из алфавита A . Пусть последовательность векторов $E_{gh} = e_1, e_2, \dots, e_c$ представляет аллофон g , за которым следует аллофон h . Здесь последовательность векторов $e_1, e_2, \dots, e_{c-c'}$ описывает аллофон g и межфонеменный переход между аллофонами g и h , а последовательность векторов $e_{c-c'+1}, \dots, e_c$ относятся к аллофону h . Будем также называть E_{gh} эталоном пары аллофонов.

Считаем b_i левую границу i -го аллофона в распознаваемом сигнале R , известной. Левую границу b_{i+1} следующего за ним $(i+1)$ -го аллофона в сигнале R получим в процессе вычисления искомой меры расхождения $d(E_{gh}, R[b_i, b_{i+1}])$.

1. Инициализация

$$1.1 \quad k_{11} := d(e_1, r_{b_i})$$

$$1.2 \quad k_{s1} := d(e_s, r_{b_i}) + k_{(s-1)1}, \quad s \in [2; c]$$

$$1.3 \quad k_{12} := k_{11} + d(e_1, r_{b_{i+1}})$$

$$1.4 \quad k_{s2} := d(e_s, r_{b_{i+1}}) + \min(k_{(s-1)2}, k_{(s-1)1}, k_{s1}), \quad s \in [2; c]$$

$$1.5 \quad f := 1; j := b_i; p := 1; q := b_i$$

2. Пока $(p \leq c) \wedge (q \leq m)$ выполнять

2.1 если $q=j+1$ то

$$2.1.1 \quad k_{1q+1} := k_{1q} + d_{1q+1}$$

$$2.1.2 \quad k_{sq+1} := d_{sq+1} + \min(k_{(s-1)q}, k_{(s-1)(q+1)}, k_{sq}), \quad s \in [2; c]$$

$$2.2 \quad f := p, j := q$$

$$2.3 \quad (p, q) := \arg \min_{(p', q') \in P_{fj}} (k_{p'q'}),$$

$$\text{где } P_{fj} = \{(p', q') : f \leq p' \leq f+1, j \leq q' \leq j+1, p' \leq n, q' \leq m, (p'=f) \vee (q'=j)\}$$

$$2.4 \quad \text{если } (p=f \wedge q=j) \text{ то } p := f+1, q := j+1$$

3. Если $(p < c) \wedge (m = q)$

3.1 то

Результат:

$b_{i+1} := m$, – номер вектора, соответствующего началу $(i+1)$ -го аллофона,

$d(E_{gh}, R[b_i, b_{i+1}]) := -1$ («отказ от распознавания»)

3.2 иначе

$$3.2.1 \quad f := p, j := q$$

$$3.2.2 \quad \text{Пока } (p > 1) \wedge (q > b_i) \geq (p \geq c-c') \text{ выполнять} \quad (p, q) := \arg \min_{(f', j') \in I} (k_{f'j'}),$$

$$\text{где } I = \{(f', j') : f-1 \leq f' \leq f, j-1 \leq j' \leq j, f' \geq 1, j' \geq b_i, f'=f-1 \vee j'=j-1\}$$

3.2.3 **Результат:**

$b_{i+1} := q+1$ – номер вектора, соответствующего началу $(i+1)$ -го аллофона,

$d(E_{gh}, R[b_i, b_{i+1}]) := k_{pq}$ – искомая мера расхождения

Рисунок 3 – Вычисление меры расхождения между эталоном i -го аллофона слова и фрагментом речевого сигнала

Для этого вычисления нам понадобится DTW-матрица K размерности $c \times m$. Заполнять её, двигаясь вправо и вниз, будем так же как в алгоритме построения пути выравнивания (рис. 2), опираясь на введённое там понятие «рабочего элемента». Заполнение будет происходить, пока рабочим элементом не станет некоторый элемент $k_{c,q}$ строки с номером c . При анализе рабочего элемента k_{p,q_h} используются элементы столбцов q_h и (q_h+1) . Поэтому алгоритм начнём с вычисления первых двух столбцов и далее при каждом изменении номера столбца рабочего элемента будем вычислять значение ещё одного столбца.

Если на некотором шаге итерации индексы рабочего элемента приняли значение (p, m) и $p < c$, делаем вывод о несоответствии распознаваемого сигнала R слову с транскрипцией T , поскольку были достигнуты границы сигнала и не достигли границы не последнего аллофона. В этом случае присвоим искомой мере расхождения значение -1 и будем говорить, что произошёл «отказ от распознавания».

По достижении индексами рабочего элемента значения (c, q) начинаем обратный проход по DTW-матрице пока текущим элементом обратного прохода не станет элемент $k_{(c-c'),q_1}$. Это и есть искомая мера расхождения, а значение (q_1+1) – левая граница $(i+1)$ -го аллофона в распознанном сигнале.

В экспериментальной части данной работы для константы c' использовано значение 2.

Для обозначения паузы, следующей за последним аллофоном слова, введём в алфавит аллофонов A специальный символ λ . Обозначим через $d'(E_{g\lambda}, R[b_l, m])$ меру расхождения между эталоном последнего аллофона $E_{g\lambda}$ слова и фрагментом $R[b_l, m]$ и будем вычислять её по алгоритму DTW.

Метод пофонемного распознавания

Метод пофонемного распознавания опирается на предлагаемые в данной работе представления словаря распознавания и функционал для вычисления меры расхождения между транскрипцией слова и распознаваемым сигналом.

Представление словаря распознавания W :

$$W(A, Et, G). \quad (12)$$

Здесь A – упорядоченное множество транскрипционных символов и символ λ , обозначающий паузу, следующую за последним аллофоном слова; $Et = \{E_{gh}\}$ – упорядоченное множество эталонов $E_{gh} = e_1, e_2, \dots, e_c$ пар аллофонов $g, h \in A$; G – множество написаний слов с ударением, по каждому из которых система распознавания строит фонетическую транскрипцию T .

Введём функционал для вычисления меры расхождения между словом словаря с транскрипцией T и распознаваемым сигналом R :

$$F(T, R) = \begin{cases} -1, \exists i = \overline{1, l-1} : d(Et_{t_i t_{i+1}}, R[b_i, b_{i+1}]), b_{i+1} \geq m \\ \sum_{i=1}^{l-1} d(Et_{t_i t_{i+1}}, R[b_i, b_{i+1}]) + d'(Et_{t_l \lambda}, R[b_l, m]), \text{ иначе} \end{cases}, \quad b_1 = 1 \quad (13)$$

Данный функционал позволяет принимать значение -1 , если происходит «отказ от распознавания» – описанный выше случай несоответствия распознаваемого сигнала транскрипции.

Результат распознавания будем вычислять по формуле:

$$res = \arg \min_j F(T_j, R) : F(T_j, R) \neq -1, \quad (14)$$

Для решения проблемы большого разброса длин эталонов автором предложено нормировать меру расхождения между эталонным и распознаваемым сигналом по длине диагонали прямоугольника, длины сторон которого равны длинам эталонного и распознаваемого сигнала. Исследование [5] показало, что этот способ нормировки меры расхождения даёт лучший результат в сравнении с другими способами решения данной проблемы. Применим его к пофонемному распознаванию.

Введем функционал $F_1(T, R)$, выполняющий нормировку меры расхождения по длине синтезированного эталона слова (последовательности эталонов аллофонов слова). Алгоритм вычисления меры расхождения $d_1(E_{gh}, R[b_i, b_{i+1}], u_i)$ с сохранением длины u_i эталона аллофона E_{gh} , будет отличаться от приведенного на рис. 3 только на этапе получения результата (рис. 3) тем, что на шагах 3.1 и 3.2.3 будет добавлена операция:

$$u_i := c - c'.$$

Аналогично поступим и с мерой расхождения последнего аллофона слова и фрагмента сигнала $d'_1(E_{gl}, R[b_l, m], u_l)$.

Функционал $F_1(T, R)$, учитывающий нормировку по длине синтезированного эталона слова, примет вид:

$$F_1(T, R) = \left\{ \begin{array}{l} \frac{\sum_{i=1}^{l-1} d_1(E_{t_{i+1}}, R[b_i, b_{i+1}], u_i) + d'_1(E_{t_l}, R[b_l, m], u_l)}{\sqrt{n^2 + m^2}}, \forall i = 1, l-1 \quad b_{i+1} < m \\ -1, \text{ иначе} \end{array} \right\}, b_1 = 1. \quad (15)$$

Здесь $n = \sum_{i=1}^l u_i$ – сумма длин эталонов u_i аллофонов, которые определяются при вычислении мер расхождения $d_1(E_{gh}, R[b_i, b_{i+1}], u_i)$ и $d'_1(E_{gl}, R[b_l, m], u_l)$.

Результат распознавания, учитывающий нормировку по длине синтезированного эталона, будем вычислять по формуле:

$$res = \arg \min_j F_1(T_j, R) : F_2(T_j, R) \neq -1. \quad (16)$$

Численное исследование

В качестве тестового набора, включающего 138 слов (рис. 4), использован речевой материал одного диктора, записанный как 8-битный сигнал с частотой дискретизации 22 050 Гц с помощью двух разных микрофонов (обозначим их «Mic» и «Mic1»), и звуковых карт («SB» и «SB1») при сходном акустическом окружении. Символ «_» в написании слов (рис. 4, 5) обозначает, что следующий за ним гласный является ударным.

Характеристики словаря выбраны так, чтобы продемонстрировать работу алгоритма DTW и предложенного метода в наиболее трудном для распознавания случае. Перечень слов тестового набора выбран исходя из следующего свойства алгоритма DTW. Он имеет тенденцию к пропуску отличающихся элементов и учёту сходных, что при отсутствии ограничений на длительность соответствующих друг другу элементов

речевого сигнала снижает значимость полученной меры расхождения. Исходя из этого наиболее трудными для распознавания по алгоритму DTW будут словари, содержащие большое количество пар фонетически близких слов. В выбранный для эксперимента словарь входят пары слов, отличающихся одним аллофоном (задыш_ать – надыш_ать, затопать – натопать, л_ай – м_ай, нас_едка – сос_едка); некоторые слова являются фрагментом других (дыш_ать – надыш_ать, задышать), другие начинаются или оканчиваются одинаковой последовательностью аллофонов (нас_едка – нас_ест, минов_ать – полинов_ать); длина слов существенно различается.

Для построения множества эталонов пар аллофонов использован набор аудио-файлов, содержащий однократное произнесение диктором слов (рис. 5). В аудио-файлах обучающего набора отмечены границы аллофонов. Этот набор записан с помощью микрофона и звуковой карты, обозначенных как «Mic» и «SB», соответственно.

_унт, алл_ель, б_оль, багат_ель, бал_ет, бал_етки, балаб_олить, балов_ать, вал_ет, воров_ать, г_оголь, г_огот, г_оль, г_опать, гал_ета, гал_етка, гогот_ун, д_ив, дыш_ать, заболел_ать, завод_ила, задыш_ать, залив_ать, засев_ать, зат_опать, кад_ило, калев_ать, карав_ай, карот_ель, кив_ать, килев_ать, кис_ель, кис_ет, ков_ать, кол_ено, кол_ет, колт_ун, л_ай, л_ен, л_ента, л_ето, л_етом, л_иф, линов_ать, лод_ыжка, лом_ать, м_ай, м_ать, м_елево, м_ель, м_ена, м_есто, м_етка, м_етод, мет_елица, мет_ель, мет_ил, минов_ать, набалов_ать, наворов_ать, надков_ать, надыш_ать, нал_ив, налив_ать, налинов_ать, налом_ать, нас_едка, нас_ест, насеv_ать, нат_опать, од_ышка, однол_етка, ол_ифа, от_ель, отков_ать, пал_етка, паралл_ель, паров_ать, пат_ент, патов_ать, пл_ед, пл_ен, пл_и, плитов_ать, побалов_ать, поворов_ать, под_и, подков_ать, подыш_ать, пол_ено, пол_ив, полив_ать, полинов_ать, полом_ать, пос_етовать, пот_опать, продыш_ать, прол_ив, пролив_ать, пролинов_ать, пролом_ать, просев_ать, прот_ест, прот_опать, с_ель, сено, с_ет, с_етка, с_етовать, своров_ать, сел_ен, синаг_ога, сов_ать, сол_ило, сос_ед, сос_едка, ст_оль, став_ать, стогов_ать, т_ент, т_ест, т_есто, т_ет-а-т_ет, т_ога, т_оль, т_опать, т_опливо, т_ополь, т_опот, т_ун, ф_и, ф_ила, ф_ифа, фас_ет, фас_етка, хал_иф, хохот_ун, целов_ать

Рисунок 4 – Слова тестового набора

_ил, -унт, б_оль, бал_етки, г_оголь, дыш_ать, забалов_ать, завод_ила, задыш_ать, зат_опать, кар_отель, кив_ать, килев_ать, кис_ет, кол_ено, кол_ет, колт_ун, л_ай, л_ента, л_етом, л_и, лод_ыжка, лом_ать, м_ай, м_елево, м_есто, мет_елица, метил, минов_ать, набалов_ать, надыш_ать, налинов_ать, налом_ать, нас_едка, нас_ест, насеv_ать, нат_опать, однол_етка, ол_ифа, паралл_ель, пл_ен, продыш_ать, прот_опать, своров_ать, сос_едка, стогов_ать, уг_одно, улом_ать, ф_ифа, хал_иф, хохот_ун, целов_ать

Рисунок 5 – Слова набора для обучения пар аллофонов

Численное исследование проведем путём сравнения частоты правильного распознавания слов: алгоритмом DTW; алгоритмом DTW с нормировкой меры расхождения между эталонным и распознаваемым сигналом по длине диагонали прямоугольника, длины сторон которого равны длинам эталонного и распознаваемого сигнала (в табл. 1 обозначено «DTW_W»); предложенным методом фонемного распознавания.

В исследовании использованы 4 тестовых набора. Символом «*» в табл. 1 отмечены наборы, в которых использовано то же оборудование (микрофон, звуковая карта), что и в обучающем наборе. Результаты численного исследования приведены в табл. 1.

Таблица 1 – Результаты численного исследования

№	Микрофон	Звуковая карта	DTW	DTW_W	Пофонемное распознавание	*
1.	Mic	SB	87	104	119	*
2.	Mic1	SB	19	31	104	
3.	Mic	SB1	16	31	95	
4.	Mic1	SB	20	38	99	

В описанном эксперименте предложенный метод пофонемного распознавания по частоте правильного распознавания на тестовых наборах превзошёл алгоритм DTW и алгоритм DTW с нормировкой меры расхождения между эталонным и распознаваемым сигналом по длине диагонали прямоугольника, длины сторон которого равны длинам эталонного и распознаваемого сигнала. Предложенный метод также более устойчив к смене звукозаписывающего оборудования (микрофона и звуковой карты).

Выводы

В работе получили дальнейшее развитие методы распознавания речи за счёт разработки метода пофонемного распознавания слов на основе алгоритма DTW.

Обоснована возможность пофонемной модификации алгоритма DTW.

Метод пофонемного распознавания опирается на предложенный функционал для вычисления меры расхождения между фонетической транскрипцией слова и распознаваемым сигналом, а также представление словаря распознавания, включающее: алфавит аллофонов; множество эталонов пар аллофонов; множество написаний слов, которое система распознавания преобразует во множество фонетических транскрипций слов.

В проведенном эксперименте метод пофонемного превзошёл алгоритм DTW по частоте правильного распознавания и показал большую устойчивость к смене звукозаписывающего оборудования.

К недостаткам работы можно отнести следующее. При обосновании возможности пофонемной модификации алгоритма DTW указано, что путь выравнивания и искомая мера расхождения сигналов не изменятся в сравнении с получаемыми по алгоритму DTW, если выдвинуть некоторые требования к эталонам пар аллофонов. Однако формально эти требования не описаны. Использованный для численного исследования эксперимент выполнен на малом количестве данных. Не рассмотрены результаты работы алгоритма DTW и предложенного метода пофонемного распознавания при использовании техники квантования векторов.

Указанные недостатки не уменьшают значение полученных результатов и являются предметом дальнейших исследований.

Список литературы

1. Jelinec F. Statistical methods for speech recognition / F. Jelinec– The MIT Press, Cambridge, Massachusetts, London, England, 1998. – P. 283.
2. Rabiner L. Fundamentals of speech recognition / L. Rabiner, B.-H. Juang. – New Jersey, 1993. – P. 507.
3. Козлов А. В. Система пофонемного распознавания отдельно произносимых слов / А. В. Козлов, Г. В. Саввина, Шелепов В. Ю. // Искусственный интеллект. – 2003. – № 1. – С. 156-165.
4. Панов М. В. Современный русский язык. Фонетика : учебник для ун-тов / Панов М. В. – М. : Высш. школа, 1979. – 256 с.
5. Шелепов В. Ю. О распознавании речи на основе межфонемных переходов / В. Ю. Шелепов, Г. В. Дорохина, А. В. Ниценко // Искусственный интеллект. – 2012. – № 1. – С.132-140.
6. Шелепов В. Ю. О некоторых вопросах, связанных с дифонным распознаванием и распознаванием слитной речи / В. Ю. Шелепов, А. В. Ниценко, Г. В. Дорохина // Искусственный интеллект. – 2013. – № 3 – С. 209-216.

7. Бурибаева А. К. Сегментация и дифонное распознавание речевых сигналов / А. К. Бурибаева, Г. В. Дорохина, А. В. Ниценко, В. Ю. Шелепов // Тр. СПИИРАН. – 2013. – Вып. 8(31). – С. 20-42.
8. Шелепов В. Ю. О некоторых подходах к проблеме компьютерного распознавания устной русской речи / В. Ю. Шелепов, О. А. Дорохин, А. В. Засыпкин, Н. А. Червин // Труды Междунар. конф. «Знание – Диалог – Решение». – Том 1. – Ялта, 1997. – С.234-240.
9. Дорохина Г. В. Анализ методов распознавания речевых команд на основе алгоритма DTW / Г. В. Дорохина // Труды шестого междисциплинарного семинара «Анализ разговорной русской речи», АР3-2012, 27-28 августа 2012, г. Санкт-Петербург. – С. 29-34.

References

1. Jelinec F. Statistical methods for speech recognition. – The MIT Press, Cambridge, Massachusetts, London, England, 1998., P. 283.
2. Rabiner L., Juang B.-H. Fundamentals of speech recognition. – New Jersey, 1993. – p. 507
3. Kozlov A.V. Isolated word recognition system based on phoneme recognition / A.V.Kozlov, G.V. Savvina, V.U. Shelepov // Artificial Intelligence. – 2003. – №1. – С. 156-165.
4. Shelepov V.Ju. On Speech Recognition Using Phoneme Transition Base / V.Ju. Shelepov, A.V. Nicenko, G.V. Dorokhina // Artificial Intelligence. — 2012. — №1. — С.132-140.
5. Nicenko A.V. On Some Questions of Diphone Recognition and Recognition of Continuous Speech / A.V. Nicenko, V.Ju. Shelepov, G.V. Dorokhina // Artificial Intelligence. – 2013. – №3 – С. 209-216.
6. Buribayeva A.K. Segmentation and diphone recognition of speech signals / A.K. Buribayeva, G.V. Dorokhina, A.V. Nitsenko, V.Ju. Shelepov // SPIIRAS Proceedings. – 2013. – Issue 8(31). – P. 20-42.
7. Panov M.V. Modern Russian. Phonetics: a textbook for University. - Higher School, Moscow 1979. – P. 256.
8. Shelepov V.Y. On some approaches to the problem of computer speech recognition of spoken Russian / V.Y. Shelepov, O.A. Dorokhin, A.V. Zasipkin, N.A. Chervin // Proceedings of the Intern. Conf. "Knowledge - Dialogue - Solution". - Volume 1 - Yalta, 1997. - S.234-240.
9. Dorokhina G.V. Analysis of methods for spoken words recognition based on the algorithm DTW // Proceedings of the Sixth Interdisciplinary Seminar "Analysis of Russian spoken speech," AR3-2012, August 27-28, 2012 in St. Petersburg. - p. 29-34.

RESUME

G. V. Dorokhina

Modification of the Algorithm DTW for Spoken Word Recognition Based on Phoneme Recognition

Background: The dominant paradigm for small vocabulary isolated speech recognition was Dynamic Time Warping (DTW). DTW gave very good results, in fact, for its field of application, better ones than HMMs did. DTW was essentially abandoned because of following problems: incorporation of language model was not natural; the problem of construction of synthetic prototypes remained unsolved; a unified recognizer statistical formulation incorporating all speech recognizer modules was never found [1 p. 53-54]. This work addresses the problem of synthetic templates – learning a small set of basic speech units smaller than recognized, and using these basic speech units to recognize units of given dictionary.

Materials and methods: Article demonstrates the possibility to modify algorithm DTW in order to recognize spoken words phoneme by phoneme. The functional for the calculation of the differences between the phonetic transcription of word and speech signal is proposed. The presentation of recognition vocabulary is developed. The proposed functional and developed presentation of recognition vocabulary form the basis of the proposed method for spoken word recognition speech recognition that uses phoneme recognition

Results: The proposed method surpassed algorithm DTW in frequency of correct recognition and showed greater resistance to change recording equipment.

Conclusion: Got the development of methods for isolated spoken word recognition through the use proposed functional for the calculation of the differences between the phonetic transcription of word and speech signal and developed presentation of recognition vocabulary. This method surpassed algorithm DTW in frequency of correct recognition and showed greater resistance to change recording equipment.

Статья поступила в редакцию 14.10.2015.