УДК 004.934.2

**Д.И. Новохрестова**

# Временная нормализация слогов алгоритмом динамической трансформации временной шкалы при оценке качества произнесения слогов в процессе речевой реабилитации

Рассмотрены проблемы временной нормализации записей слогов перед сравнительной оценкой качества произнесения. Для временной нормализации использован алгоритм динамической трансформации временной шкалы (dynamic time warping, DTW). Проведено сравнение количественных оценок качества произнесения слогов, подсчитанных на основе DTW-расстояния, метрики Минковского со значением параметра, равным 3, и коэффициента корреляции. Проведено сравнение оценок, полученных с применением нормализации по интенсивности и сглаживания временных последовательностей значений амплитуд звуковых файлов перед применением DTW. Сделан вывод о применимости DTW для временной нормализации слогов для оценки разборчивости в процессе речевой реабилитации. В качестве итоговой оценки качества произнесения выбрано DTW-расстояние, получаемое на основе алгоритма без предварительной нормализации и сглаживания.

**Ключевые слова:** оценка качества речи, качество произнесения слогов, временная нормализация, динамическая трансформация, временная шкала.
**doi:** 10.21293/1818-0442-2017-20-4-142-145

Ежегодно в России регистрируется более 25 000 новых случаев раковых заболеваний органов речеобразующего аппарата, общее же количество пациентов с такой локализацией – более 100 000 [1, 2]. Зачастую в ходе лечения прибегают к хирургическому вмешательству, а после операции встает вопрос о необходимости речевой реабилитации. В настоящий момент оценка качества речи в процессе реабилитации оценивается субъективным методом, а именно экспертной оценкой согласно ГОСТ Р 58040–95 [3]. Ввиду субъективности оценки невозможно точно оценить динамику восстановления речи, а также своевременно заметить возможное ухудшение качества речи из-за неэффективности занятий для данного пациента. Поэтому возникла необходимость в автоматизации процесса оценки качества речи с минимизацией участия специалиста-логопеда в нем. Конечная цель, достигаемая путем автоматизации процесса, – сокращение времени реабилитации. Также была поставлена задача получения оценки в течение времени, приближенного к реальному, в идеальном варианте – получение оценки в режиме реального времени при записи сеанса оценки качества произношения слогов для реализации биологической обратной связи.

Ранее была предложена математическая модель и реализованный в Matlab модуль для оценки качества произношения слогов на основе сравнения спектрограмм Фурье [4]. Существенным минусом предложенного подхода является не автоматизированный на настоящее время этап по временной нормализации слогов, а именно сегментация слогов на фонемы. Существующие в настоящий момент алгоритмы сегментации либо имеют большую ошибку точности определения границ фонемы, либо не могут выполняться в режиме реального времени. Поэтому было принято решение о применении иного метода временной нормализации – алгоритма динамической трансформации временной шкалы (dynamic time warping – DTW). Хоть данный алгоритм и обладает некоторыми недостатками (отсутствие возможности оценки только конкретной проблемной фонемы), но на данный момент является наилучшим решением задачи временной нормализации в процессе оценки качества произношения слогов.

**Алгоритм DTW**

DTW предназначен для трансформирования временных последовательностей на основе поиска наибольшего подобия, к которым и относятся значения амплитуд звуковых файлов [5]. Этот алгоритм был выбран для реализации ввиду простоты его реализации, а также квадратичной сложности классического алгоритма, что является важным параметров из-за необходимости оценки в режиме реального времени для организации биологической обратной связи.

DTW основан на построении матриц расстояний между всеми элементами первой последовательности и всеми элементами второй последовательности (матрица $d$ размера $m \times n$, где $d_{i,j}$ – расстояние между точками $p_i$ и $q_i$ последовательностей $P$ и $Q$ соответственно, $m$ – длина последовательности $P$, $n$ – длина последовательности $Q$) и построении матрицы деформаций на основе матрицы расстояний (матрица $D$ размера $m \times n$, где $D_{i,j}$ определяется согласно (1)). Расстояние $d_{i,j}$ в DTW принято считать на основе расстояния городских кварталов (2) или евклидова расстояния (3) [6]:

$$D_{i,j} = d_{i,j} + \min\left(D_{i-1,j},\ D_{i,j-1},\ D_{i-1,j-1}\right), \quad (1)$$

$$d_{i,j} = d(p_i, q_j) = \left|p_i - q_j\right|, \quad (2)$$

$$d_{i,j} = d(p_i, q_j) = (p_i - q_j)^2, \quad (3)$$

где $p_i$ и $q_i$ – точки последовательностей $P$ и $Q$ соответственно.

Трансформированные последовательности получаются путем выбора элементов каждой из после-

довательности согласно индексам элементов, принадлежащих оптимальному пути. Также преимуществом алгоритма является то, что по итогам выполнения алгоритма уже подсчитана мера разности между временными последовательностями – DTW-расстояние. Оно представляет собой стоимость пути, основанного на оптимальном пути трансформации. Значение DTW-расстояния равно значению элемента $D_{n,m}$ матрицы $D$. Принципиальная блок-схема алгоритма приведена на рис. 1.

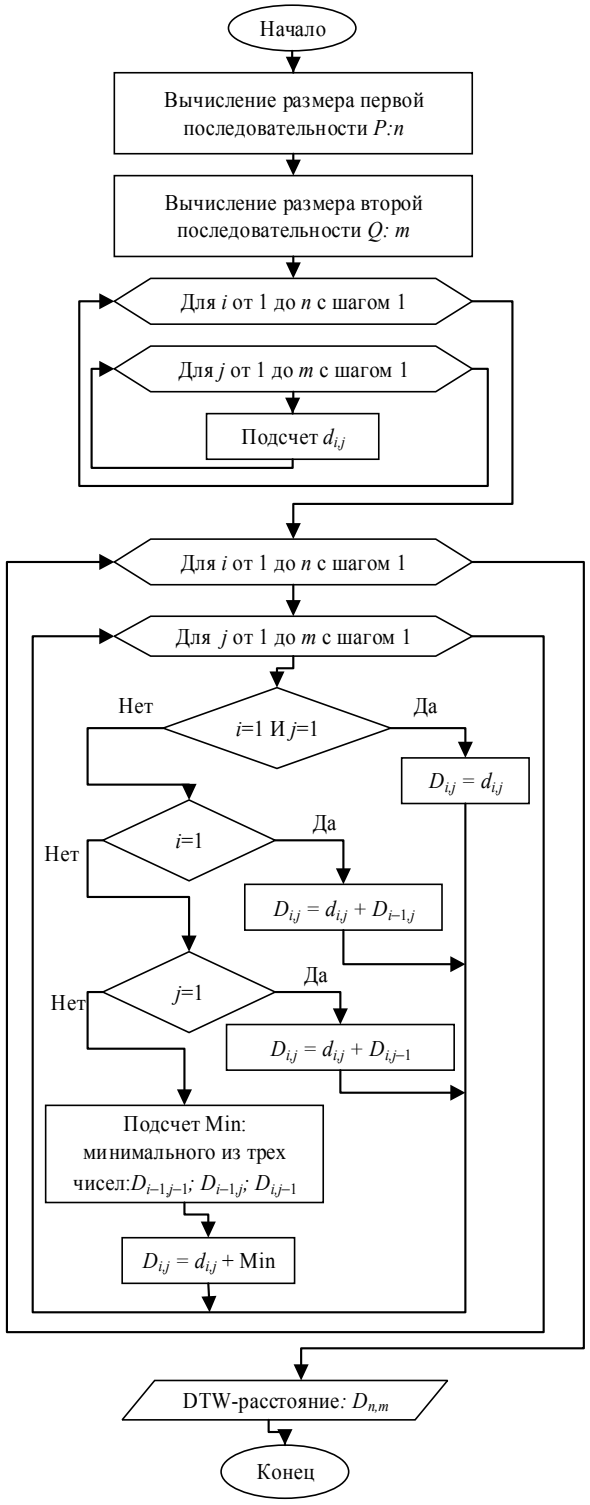


Рис. 1. Блок-схема DTW с подсчетом DTW-расстояния

**Количественные оценки**

Помимо DTW-расстояния, были подсчитаны расстояния между трансформированными последовательностями на основе метрики Минковского (4) [6] и коэффициента корреляции (5) [7]. Применимость коэффициента корреляции для сравнительной оценки качества произношения слогов рассмотрена в [8].

$$d(p,q)=\sqrt[t]{\sum_{k=1}^{n}\left|p_k-q_k\right|^t}, \tag{4}$$

где $p$ и $q$ – точки трансформированных временных последовательностей; $n$ – длина трансформированной последовательности; $t$ = 3 – параметр метрики Минковского.

$$r_{pq}=\frac{\sum_{i=1}^{n}\left(p_i-\overline{p}\right)\left(q_i-\overline{q}\right)}{\sqrt{\sum_{i=1}^{n}\left(p_i-\overline{p}\right)^2\sum_{k=1}^{n}\left(q_i-\overline{q}\right)^2}}, \tag{5}$$

где $r_{pq}$ – коэффициент корреляции; $\overline{p}$ и $\overline{q}$ – выборочные средние трансформированных числовых последовательностей $P^*$ и $Q^*$ соответственно.

При условии наличия нескольких метрик, которые могут быть использованы как мера различия между временными последовательностями, есть необходимость в проведении сравнения оценок, получаемых на основе этих метрик. Для этого на основе списка слогов с наиболее подверженными изменениям фонемами [4] было произведено 3 сеанса записи: два сеанса с нормальным произношением слогов и третий сеанс с произношением слогов без использования языка (искаженное произношение с имитацией изменений, характерных для речи пациентов с раком языка, перенесших операцию).

Были подсчитаны меры различия для пар слогов норма1–норма2, норма1–искажение, норма2–искажение, где норма1 – запись слога первого сеанса, норма2 – запись этого же слога из второго сеанса, искажение – запись этого же слога из третьего сеанса. Если говорить о DTW-расстоянии и метрике Минковского, количественная оценка пары норма1–норма2 должно быть меньше, чем оценка норма1–искажение и норма2–искажение.

Для коэффициента корреляции смотрелась близость коэффициента к 1, соответственно коэффициент для пары норма1–норма2 должен быть ближе к 1, чем коэффициент норма1–искажение и коэффициент норма2–искажение. В табл. 1 приведены результаты работы алгоритма, указаны средние оценки каждой пары сеансов, найденные как среднее между соответствующими парами слогов этих сеансов, относительные оценки сеансов (как отношение среднего оценок норма1–искажение и норма2–искажение к средней оценке норма1–норма2), количество ошибок для каждой меры различия (суммарное количество ошибок, количество ошибок для пар норма1–искажение, количество ошибок для пар норма2–искажение), время работы алгоритма (среднее время

на одну пару слогов при подсчете оценок для трех сеансов).

Таблица 1

**Сравнение результаты работы алгоритмов оценки на основе различных мер различия**

| | | Мера различия | | |
|---|---|---|---|---|
| | | DTW-расстояние | Метрика Минковского | Коэффициент корреляции |
| Оценка сеанса | Норма1–норма2 | 1,405 | 0,137 | 0,401 |
| | Норма1–искажение | 1,480 | 0,141 | 0,365 |
| | Норма2–искажение | 2,248 | 0,214 | 0,413 |
| | Общая | 1,864 | 0,178 | 0,389 |
| Относительная оценка сеанса | | 0,754 | 0,772 | 0,970 |
| Количество ошибок | Норма1–искажение | 32 | 40 | 52 |
| | Норма2–искажение | 2 | 16 | 46 |
| | В обеих парах | 1 | 12 | 33 |
| Время выполнения, с | | 3,02 | 3,30 | 5,32 |

Сравнение показало, что наименьшим количеством ошибок обладают оценки, полученные на основе DTW-расстояния, а также время, затрачиваемое на его подсчет, минимальное среди представленных.

**Нормализация и сглаживание временных последовательностей**

При анализе пар слогов, для которых были допущены ошибки при подсчете количественной оценки, возникло предположение о чувствительности DTW к различиям в интенсивности голоса на записи, а также к наличию случайных кратковременных шумов. Поэтому была рассмотрена возможность нормализации или сглаживания временной последовательности перед применением DTW-алгоритма. Были сравнены результаты, полученные на основе последовательностей без нормализации и сглаживания, последовательностей с нормализацией по интенсивности (приведение интенсивности звука к константе), сглаженных на основе простого скользящего среднего (6) последовательностях, а также нормализованных по интенсивности сглаженных на основе простого скользящего среднего последовательностей.

$$\bar{y}(k)=\frac{1}{n}\sum_{t=k-\frac{n-1}{2}}^{t=k+\frac{n-1}{2}} y(t), \qquad (6)$$

где $n$ – размер окна сглаживания (обязательно нечетное число); $y(t)$ – $t$-й элемент временного ряда; $\bar{y}(k)$ – $k$-й элемент сглаженного временного ряда.

Применение сглаживания только простой скользящей средней на данном этапе исследований обосновано тем, что данное сглаживание обладает минимальной сложностью алгоритма (линейной сложностью). Его применение может показать необходимость сглаживания последовательностей, а при наличии необходимости сглаживания подбор алгоритма сглаживания – отдельная задача.

Результаты приведены в табл. 2, подсчитана средняя оценка сеанса, относительная оценка сеанса, количество ошибок, среднее время выполнения одного расчета. Размер окна сглаживания $n$ в скользящей средней равен 13.

Таблица 2

**Сравнение результатов работы алгоритма DTW с нормализацией и сглаживанием числовых последовательностей**

| | | Алгоритм | | | |
|---|---|---|---|---|---|
| | | DTW | DTW с нормализацией по интенсивности | DTW со сглаживанием | DTW с нормализацией по интенсивности и сглаживанием |
| Оценка сеанса | Норма1–норма2 | 1,405 | 4,410 | 0,225 | 0,402 |
| | Норма1–искажение | 1,480 | 4,878 | 0,259 | 0,445 |
| | Норма2–искажение | 2,248 | 4,723 | 0,232 | 0,430 |
| | Общая | 1,864 | 4,800 | 0,246 | 0,438 |
| Относительная оценка сеанса | | 0,738 | 0,754 | 0,919 | 0,918 |
| Количество ошибок | Норма1–искажение | 32 | 29 | 33 | 29 |
| | Норма2–искажение | 2 | 37 | 42 | 36 |
| | В обеих парах | 1 | 19 | 24 | 19 |
| Время выполнения, с | | 3,02 | 3,78 | 3,26 | 3,80 |

Применение нормализации по интенсивности незначительно уменьшило количество ошибок в парах норма1–искажение по сравнению с оценками по DTW без нормализации, однако сильно увеличилось количество ошибок в парах норма2–искажение, а также в 19 раз возросло количество ошибок в обеих парах, в связи с чем было принято решение о реализации алгоритма DTW для использования в процессе автоматизированной оценки качества произношения слогов без предварительной нормализации и сглаживания.

**Заключение**

В работе рассмотрено использование алгоритма DTW для временной трансформации записей для подсчета количественной оценки качества произношения слогов в процессе речевой реабилитации. Проведено сравнение оценок, получаемых на основе предлагаемых мер различия, для использования в качестве количественной оценки предложено DTW-расстояние, обладающие наименьшим количеством ошибок для сравниваемых пар. Применение нормализации по интенсивности и сглаживания к временным последовательностям перед применением DTW показывает результат хуже, чем применение DTW к первоначальным последовательностям.

Время подсчета оценки для одной пары записей в среднем составляет 3 с, что, несомненно, быстрее, чем ручное проставление оценок, однако для реализации биологической обратной связи появляется

необходимость сокращения времени подсчета, что может потребовать либо оптимизацию программного кода, либо переработку самого алгоритма (как вариант, применение не классического DTW, а ускоренного DTW).

Работа выполнена при поддержке Российского научного фонда, проект «Восстановление речевой функции с использованием технических методов и математического моделирования у больных раком полости рта и ротоглотки после хирургического лечения», № 1615-00038.

*Литература*

1. Злокачественные новообразования в России в 2016 г. (заболеваемость и смертность) / Под ред. А.Д. Каприна, В.В. Старинского, Г.В. Петровой. – М.: МНИОИ им. П.А. Герцена, 2018. – 250 с.

2. Состояние онкологической помощи населению России в 2016 году / под ред. А.Д. Каприна, В.В. Старинского, Г.В. Петровой. – М.: МНИОИ им. П.А. Герцена, 2017. – 236 с.

3. ГОСТ Р 50840–95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. – М.: ИПК Изд-во стандартов, 1996. – 234 с.

4. Model of system quality assessment pronouncing phonemes / Е.Ю. Костюченко, Д.И. Игнатьева, Р.В. Мещеряков и др. // Dynamics of Systems, Mechanisms and Machines (Dynamics). – 2016 [Электронный ресурс]. – Режим доступа: http://ieeexplore.ieee.org/document/7819016/, свободный (дата обращения: 26.11.2017).

5. Романенко А.А. Выравнивание временных рядов: прогнозирование с использованием DTW / А.А. Романенко // Машинное обучение и анализ данных. – 2001. – № 1 [Электронный ресурс]. – Режим доступа: http://jmlda.org/papers/doc/2011/no1/Romanenko2011Dynamic.pdf (дата обращения: 26.11.2017).

6. Теслер Г.С. Метрики и нормы в иерархии категориальных семантик и функций // Математические машины и системы. – 2005. – № 2 [Электронный ресурс]. – Режим доступа: http://elektronika.vk.ru/4(28)2008/4.html, свободный (дата обращения: 26.11.2017).

7. Гмурман В.Е. Теория вероятностей и математическая статистика: учеб. пособие для вузов. – 10-е изд., стереот. – М.: Высшая школа, 2004. – 479 с.

8. Correlation normalization of syllables and comparative evaluation of pronunciation quality in speech rehabilitation / E. Kostyuchenko, R. Meshcheryakov, D. Ignatieva et al. // 19th International conference on speech and computer SPECOM 2017, Lecture Notes in Computer Science. Springer, Cham. – 2017. – Vol. 10458. – P. 262–271.

9. Новикова Н.В. Прогнозирование национальной экономики: учеб.-метод. пособие / Н.В. Новикова, О.Г. Поздеева. – Екатеринбург: Урал. гос. экон. ун-т, 2007. – 137 с.

---

**Новохрестова Дарья Игоревна**
Студентка каф. безопасности информационных систем ТУСУРа
Тел.: +7 (382-2) 70-15-29 (внутр. 29-66)
Эл. почта: ndi@fb.tusur.ru

Novokhrestova D.I.

**Time normalization of syllables with the dynamic time warping algorithm in assessing of syllables pronunciation quality when speaking**

The article is devoted to the problem of time normalization of syllable records before the comparative evaluation of the pronunciation quality. The dynamic time warping (DTW) algorithm is used for time normalization. A comparison is made of quantitative estimates of the syllables pronunciation quality. Estimates are calculated on the basis of the DTW-distance, the Minkowski metric with a parameter value of 3 and the correlation coefficient. A conclusion is made about applicability of DTW for time normalization of syllables for evaluating intelligibility in the process of speech rehabilitation. As the final evaluation of the pronunciation quality the DTW-distance was chosen obtained on the basis of the algorithm without preliminary normalization and smoothing.

**Keywords:** speech quality estimation, syllables pronunciation quality, time normalization, DTW.