Proceedings of IAM, V.13, N.2, 2024, pp.210-219
DOI: 10.30546/2225-0530.2024.13.2.026

КОМПЛЕКСНЫЙ АЛГОРИТМ РАСПОЗНАВАНИЯ ГОЛОСОВОГО СИГНАЛА НА ПРИМЕРЕ ОЗВУЧЕННЫХ СЛОВ ИЗ АЗЕРБАЙДЖАНСКОГО ЛЕКСИКОНА

Э.Р. Исмаилов

Институт Систем Управления Министерства Науки и Образования Азербайджанской Республики, Баку, Азербайджан
e-mail: elchinisri@gmail.com

Абстракт. Наряду с алгоритмами компьютерного зрения для обработки фото и видео информации, а также техники работы с естественными языками для анализа текстов, работа с аудио информацией также является наиболее востребованной процедурой для ведения бизнес-аналитики. В статье рассматривается задача распознавания речевого сигнала на примере базы аудиоданных, формируемой на основе слов, воспроизведенных на азербайджанском языке. В предлагаемом подходе звуковой сигнал рассматривается как одномерное представление колебаний звуковой волны с определенной частотой семплирования. Для реализации поставленной задачи применяются евклидовый метод распознавания, а также методы DTW и DDTW. Эмпирический анализ результатов распознавания голосовых воспроизведений слов на азербайджанском языке выявил качественные преимущества метода DTW над остальными.

Ключевые слова: Голосовой сигнал, база аудиоданных, евклидова метрика, метод распознавания.

AMS Subject Classification: 94-10.

1. Введение

Темп жизни стремительно растет, объем знаний ежегодно удваивается, свободного времени становится все меньше, и неудивительно, что часть повседневных задач люди стараются делегировать машинам. Это актуально не столько для бытовой сферы, сколько для бизнеса, где виртуальные ассистенты оперативно выполняют скучную и рутинную работу, оставляя сотрудникам больше свободного времени на креативные и стратегические задачи. Виртуальный помощник – это цифровая программа, приложение или сервис на базе искусственного интеллекта (ИИ), использующий цифровую звукозапись, в качестве технологии преобразования аналогового звука в цифровой с целью сохранения его на физическом носителе для возможности последующего воспроизведения записанного сигнала. Представление аудиоданных в цифровом виде позволяет очень эффективно изменять исходный материал при помощи специальных устройств или компьютерных программ – звуковых редакторов, что нашло широкое применение в промышленности, медиа-индустрии и быту.

Принято считать, что цифровая звукозапись – это представление звука, записанного или преобразованного в цифровой сигнал. В процессе аналого-цифрового преобразования амплитуды аналоговой звуковой волны фиксируются с заданной частотой дискретизации и битовой глубиной и преобразуются в аудиоданные, которые уже могут редактироваться компьютерной программой. Аналого-цифровое преобразование (или иначе, квантование) очень похоже на сьёмку видеокамерой, которая восстанавливает непрерывный момент времени, захватывая тысячи последовательных изображений в секунду, называемых кадрами. Чем выше частота кадров, тем отчётливее выглядит запись. В цифровом аудио аналого-цифровой преобразователь захватывает тысячи аудиосэмплов (аудио отсчётов) в секунду с указанной частотой дискретизации и битовой глубиной для восстановления исходного сигнала. Чем выше частота дискретизации и битовая глубина, тем выше разрешение звука.

Целью статьи является разработка комплексного алгоритма, способного с высокой точностью распознавать слова на азербайджанском языке в голосовом исполнении путем цифровизации звука, как результата преобразования аналогового сигнала звукового диапазона в цифровой аудио формат. В качестве инструментов распознавания голосового сигнала применяются Евклидовый метод распознавания, а также хорошо зарекомендовавшие себя в контекстной области методы распознавания DTW (Dynamic Time Warping) и DDTW (Derivative Dynamic Time Warping) [1, 2].

2. Постановка задачи

Несколько слов, например, «книга», «тетрадь» и «карандаш» воспроизводятся человеком на азербайджанском языке и через звуковое устройство преобразуются в аналоговые сигналы. Далее, путем квантования эти сигналы трансформируются в соответствующие цифровые сигналы s_1, s_2 и s_3, которые и формируют базу аудиоданных, представленную на Рис. 1.

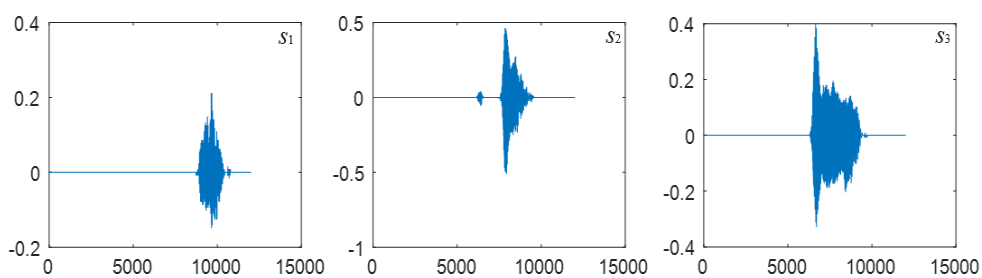

Рис. 1. База аудиоданных, включающая цифровые сигналы s_1, s_2 и s_3

Предположим, что одно из указанных слов воспроизводится тем или иным человеком, которое после соответствующих преобразований представляется в виде цифрового сигнала s, отраженного на Рис. 2.

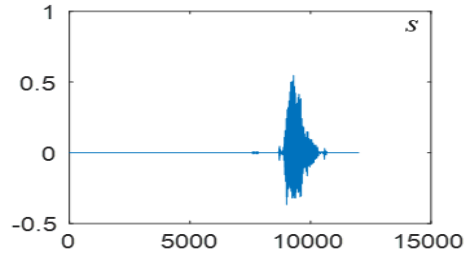

Рис. 2. Распознаваемое слово «книга», представленное в виде цифрового сигнала *s*

На базе этого тривиального примера необходимо сформулировать подход к распознаванию голосовых сигналов, предполагающий применение двух и более методов распознавания [3, 4, 5].

3. Критерии оценки близости сигналов

На предварительной стадии распознавания сигналов, как правило, выявляются основные признаки распознавания и на их основе производится выбор соответствующей нормы расстояния. Далее осуществляется процедура распознавания посредством сравнения распознаваемых сигналов с эталоном путем вычисления попарных расстояний между ними на основе выбранной метрики. Выбор признаков распознавания зависит от характера решаемой задачи (семейства распознаваемых сигналов) и применяемого подхода. Тем не менее, во всех случаях в качестве базовых норм расстояния между сигналами используется Евклидова метрика.

Евклидовый метод распознавания. В качестве признаков распознавания выбираются значения точек отсчётов. В частности, если для двух произвольных сигналов x и y отсчётами являются соответственно точки a_i и b_i ($i = 0, 1, \ldots, N$), тогда в качестве нормы расстояния между ними выбирается евклидова метрика в виде

$$D_1(x,y) = \sqrt{\sum_{i=1}^{N}(a_i - b_i)^2}\ . \qquad (1)$$

Метод распознавания DTW [6]. Этот метод общеизвестен [7, 8, 9] и уже давно применяется. Тем не менее, для дальнейшего повествования опишем суть метода на тривиальном примере сравнения двух числовых последовательностей $\{f_1, f_2, \ldots, f_n\}$ и $\{g_1, g_2, \ldots, g_m\}$ с длинами n и m, соответственно (см. Рис. 3). Алгоритм начинается с расчета локальных отклонений между компонентами этих последовательностей. Самым распространенным является метод, рассчитывающий абсолютное отклонение между значениями двух компонент (евклидово расстояние). В результате формируется матрица размера $n \times m$, состоящая из квадратичных отклонений

вида $d_{ij} = (f_i - g_j)^2$, i=1÷n, j=1÷m, и вычисляется минимальное расстояние DTW (f_i, g_j) с применением следующих равенств:

$$\begin{cases} DTW(f_i, g_j)^2 = d_{ij} + \min\{DTW(f_i, g_{(j-1)})^2, DTW(f_{(i-1)}, g_j)^2, DTW(f_{(i-1)}, g_{(j-1)})^2\}, \\ DTW(f_1, g_1)^2 = d_{11}. \end{cases}$$

По результатам итеративного вычисления $DTW(f_i, g_j)$ норма расстояния между сигналами формируется в виде

$$D_2 = \sqrt{DTW(f_n, g_n)} \tag{2}$$

При этом, применение метода *DTW* подразумевает выполнение следующих условий:

- *Монотонность* – оба индекса *i* и *j* последовательно возрастают.
- *Непрерывность* – за один шаг индексы *i* и *j* увеличиваются не более чем на единицу.
- Последовательное построение «путей» начинается в левом нижнем и заканчивается в правом верхнем углу.

Алгоритм DTW применяется с «ограничением» и «без ограничения» на размер так называемого «окна», размер которого *w* определяет число разрешенных отсчетов, позволяющих проводить сравнение компонентов сигналов как справа, так и слева. При этом общее число отсчетов составляет 2*w*+1, где процедура сравнения f_i и g_j по *i*-ой точке отсчета у первого и *j*-ой точке отсчета у второго должна соответствовать неравенству $|i - j| \le w$.

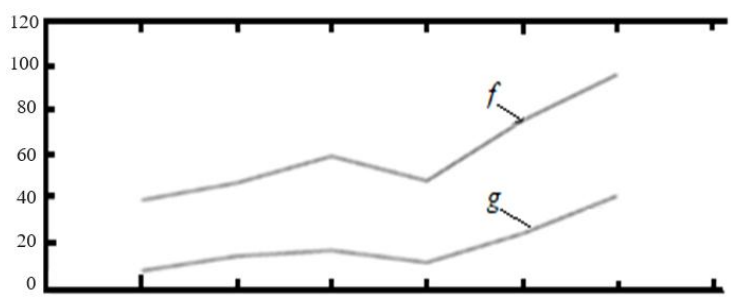

Рис. 3. Числовые последовательности сигналов f_i и g_j

Метод распознавания DDTW [10]. В качестве признаков распознавания выбираются значения первых производных в точках отсчётов. В дискретном случае, в качестве производной первого порядка принимается выражение [11]: $\dot{a}(i) = [a(i) - a(i-1)]/T$, где $a(i) = a(iT)$, i = 0, 1, …, N; T – период дискретизации аналогового сигнала a. В частности, если для двух произвольных сигналов x и y отсчётами являются соответственно значения первых производных p_i и q_i (i = 0, 1, …, N), тогда в качестве нормы

расстояния между ними выбирается евклидово расстояние

$$D_3(x,y)=\sqrt{\sum_{i=1}^{N}(p_i-q_i)^2}\ . \qquad (3)$$

4. Распознавание голосового сигнала с применением критериев оценки близости

Сравнения распознаваемого сигнала s с сигналами из сформированной базы аудиоданных осуществлены для разных размеров окон w. Так, для случая w=5, результаты попарного сравнения сигналов из семейства $S = \{s_1, s_2, s_3, s\}$ с применением метрик (1), (2) и (3) сведены в Таблицы 1-3.

Таблица 1. Результаты попарного сравнения сигналов из семейства S с применением (1)

	s_1	s_2	s_3	s
s_1	0	5.97193959279563	5.61274163420549	4.62011460727708
s_2	5.97193959279563	0	7.97087517617851	7.17778725533242
s_3	5.61274163420549	7.97087517617851	0	6.87520862609592
s	4.62011460727708	7.17778725533242	6.87520862609592	0

Таблица 2. Результаты попарного сравнения сигналов из семейства S с применением (2)

	s_1	s_2	s_3	s
s_1	0	5.92235424347077	5.54185512289206	3.40050224633711
s_2	5.92235424347077	0	6.79303950689141	7.03453916123420
s_3	5.54185512289206	6.79303950689141	0	6.72446613674415
s	3.40050224633711	7.03453916123420	6.72446613674415	0

Таблица 3. Результаты попарного сравнения сигналов из семейства S с применением (3)

	s_1	s_2	s_3	s
s_1	0	2.732616622125353	1.810623776346889	2.644906269568866
s_2	2.732616622125353	0	3.133034106096033	3.622818915503668
s_3	1.8106237763468	3.1330341060960	0	2.9952966059828

	9	3		7
s	2.6449062695688 6	3.6228189155036 8	2.9952966059828 7	0

При выборе окна размером w=25 результаты аналогического попарного сравнения сигналов помещены в Таблицы 4-6.

Таблица 4. Результаты попарного сравнения сигналов из семейства S с применением (1)

	s_1	s_2	s_3	s
s_1	0	5.9719395927956 3	5.6127416342054 9	4.7633511481053 2
s_2	5.9719395927956 3	0	7.9708751761785 1	7.1674269635165 3
s_3	5.6127416342054 9	7.9708751761785 1	0	6.9335838468239 1
s	4.7633511481053 2	7.1674269635165 3	6.9335838468239 1	0

Таблица 5. Результаты попарного сравнения сигналов из семейства S с применением (2)

	s_1	s_2	s_3	s
s_1	0	5.8628062869750 1	5.4312722507455 6	4.0047883936763 8
s_2	5.8628062869750 1	0	5.7053131504354 1	6.6251151596712 8
s_3	5.4312722507455 6	5.7053131504354 1	0	5.9802735618866 1
s	4.0047883936763 8	6.6251151596712 8	5.9802735618866 1	0

Таблица 6. Результаты попарного сравнения сигналов из семейства S с применением (3)

	s_1	s_2	s_3	s
s_1	0	2.7326166221253 3	1.8106237763468 9	2.4852577416599 0
s_2	2.7326166221253 3	0	3.1330341060960 3	3.5658275389589 0
s_3	1.8106237763468 9	3.1330341060960 3	0	2.9260941504912 4
s	2.4852577416599 0	3.5658275389589 0	2.9260941504912 4	0

Как видно из абсолютно всех результатов попарного сравнения, цифровой сигнал s, отражающий голосовой сигнал «книга», воспроизведенного на азербайджанском языке, распознается всеми тремя методами как сигнал s_1. С точки зрения близости, кажется, что наиболее точным является амплитудный метод распознавания, и далее – методы DDTW и DTW. Однако качество распознавания определяется величиной отношения $(s_j, s)/(s_1, s)$, $(j = 2, 3)$, где (s_j, s) – оценка близости распознаваемого сигнала s с j-ым сигналом s_j; (s_1, s) – оценка близости распознаваемого сигнала s с его аналогом s_1 из базы аудиоданных. Чем больше это отношение, тем качественнее распознавание, которое демонстрирует тот или иной метод.

В частности, для случая w=5 (см. Таблицы 1-3) имеем:

- для амплитудного метода $\frac{(s_2, s)}{(s_1, s)} = \frac{5.97193959279563}{4.62011460727708} = 1.292$;
- для метода DTW $\frac{(s_2, s)}{(s_1, s)} = \frac{5.92235424347077}{3.40050224633711} = 1.741$;
- для метода DDTW $\frac{(s_2, s)}{(s_1, s)} = \frac{2.73261662212533}{2.64490626956886} = 1.034$.

Для случая w=25 (см. Таблицы 4-6) соответственно имеем:

- для амплитудного метода $\frac{(s_2, s)}{(s_1, s)} = \frac{5.97193959279563}{4.76335114810532} = 1.254$;
- для метода DTW $\frac{(s_2, s)}{(s_1, s)} = \frac{5.86280628697501}{4.00478839367638} = 1.465$;
- для метода DDTW $\frac{(s_2, s)}{(s_1, s)} = \frac{2.73261662212533}{2.48525774165990} = 1.1008$.

Из приведенных расчетов видно, что в обоих случаях размера «окна» наибольшая величина отношения $\frac{(s_2, s)}{(s_1, s)}$ достигается при применении метода DTW, что с точки зрения качества, выгодно отличает этот метод от остальных.

5. Заключение

На тривиальном примере голосовых воспроизведений трех слов на азербайджанском языке в статье сформулирован подход к распознаванию голосовых сигналов, основанный на комбинированном применении пока еще трех методов распознавания. Перспективная система, комбинирующая в себе различные методы распознавания, способна будет реагировать на голосовые команды и выполнять задачи на основе действий пользователя [12, 13]. Это всего лишь один пример того, как один из инструментов ИИ может быть интегрирован в обычные устройства [14, 15], чтобы сделать их более интуитивно понятными и способными взаимодействовать с гражданами Азербайджана на естественном языке.

Приведенные в статье расчеты получены с применением авторского программного обеспечения к.ф.-м.н, доцента А.Б. Керимова, которое применялось в [1, 2, 5] и в написании ряда других рецензируемых статей под руководством д.т.н., профессора Рзаева Р.Р.

Литература

1. Afouras T., Chung J.S., Senior A., Vinyals O., Zisserman A. Deep audio-visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell., V.44, (2018), pp.8717-8727.
2. Elmir B.S., Abdeslam Y.D. A study on automatic speech recognition, Journal of Information Technology Review, V.10, (2019), pp.77-85.
3. Geler Z., Kurbalija V., Ivanović M., Radovanović M., Dai W. Dynamic Time Warping: Itakura vs Sakoe-Chiba, IEEE International Symposium On Innovations in Intelligent SysTems and Applications (INISTA), (2019). https://ieeexplore.ieee.org/document/8778300
4. Haridas A.V., Marimuthu R., Sivakumar V.G. A critical review and analysis on techniques of speech recognition: the road ahead, Int. J. Knowl. Base. Intell. Eng. Syst., V.22, (2018), pp.39-57.
5. Itakura F. Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, V.23, N.1, (1975), pp.67-72.
6. Jiang Sh., Chen Z. Application of dynamic time warping optimization algorithm in speech recognition of machine translation, V.9, N.11, (2023), https://doi.org/10.1016/j.heliyon.2023.e21625
7. Keogh E.J., Pazzani M.J. Derivative Dynamic Time Warping, Proceedings of the 2001 SIAM International Conference on Data Mining, https://doi.org/10.1137/1.9781611972719.1
8. Kerimov A.B. Accuracy comparison of signal recognition methods on the example of a family of successively horizontally displaced curves, Informatics and Control Problems, V.42, N.2, (2022), pp.80-91.
9. Kerimov A.B. Comparison of some signal recognition methods for their adequacy, Proceedings of the 8th International Conference on Control and Optimization with Industrial Applications, V.I, 24-26 August, Baku, Azerbaijan, (2022).
10. Linh L.H., Hai N.T., Thuyen N.V., Mai T.T., Toi V.V. MFCC-DTW Algorithm for Speech Recognition in an Intelligent Wheelchair, 5th International Conference on Biomedical Engineering in Vietnam, (2014), pp. 417-421.
11. Novozhilov B.M. Calculation of the derivative of an analog signal in a programmable logic controller, Aerospace scientific journal of Moscow State Technical University. N.E. Bauman, Electron. Magazine, N.4, (2016), pp.1-12. (In Russian).

12. Rajeev R., Abhishek T. Analysis of feature extraction techniques for speech recognition system, Int. J. Innovative Technol. Explor. Eng., V.8, (2019), pp.197-200.
13. Rzayev R.R., Kerimov A.B. Signal recognition using weighted additive convolution of evaluation criteria, The Springer Series "Lecture Notes in Networks and Systems", V.2, (2023), pp.407-416, https://doi.org/10.1007/978-3-031-39777-6_49
14. Sakoe H., Chiba S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, V.ASSP-26, N.1, (1978), pp.43-49.
15. Zhi-Qiang U., Jia-Qi Z., Xin W., Zi-Wei L., Yong L. Improved algorithm of DTW in speech recognition, IOP Conference Series: Materials Science and Engineering, V.563, N.5, (2024), DOI 10.1088/1757-899X/563/5/052072.

COMPLEX ALGORITHM FOR VOICE SIGNAL RECOGNITION USING THE EXAMPLE OF VOICED WORDS FROM THE AZERBAIJANI LEXICON

E.R. Ismailov
Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan, Baku, Azerbaijan
e-mail: elchinisri@gmail.com

Abstract. Along with computer vision algorithms for processing photo and video information, as well as natural language techniques for text analysis, working with audio information is also the most actual procedure for conducting business analytics. The article considers the problem of speech signal recognition using the example of the audio database formed on the basis of words reproduced in the Azerbaijani language. In the proposed approach, the sound signal is considered as a one-dimensional representation of sound wave oscillations with a certain sampling frequency. To implement this problem, the Euclidean recognition method, as well as the DTW and DDTW methods, are used. An empirical analysis of the results of recognizing voice reproductions of words in the Azerbaijani language revealed the qualitative advantages of the DTW method over the others.

Keywords: Voice signal, audio database, Euclidean metric, recognition method.

References

1. Afouras T., Chung J.S., Senior A., Vinyals O., Zisserman A. Deep audio-visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell., V.44, (2018), pp.8717-8727.
2. Elmir B.S., Abdeslam Y.D. A study on automatic speech recognition, Journal of Information Technology Review, V.10, (2019), pp.77-85.

3. Geler Z., Kurbalija V., Ivanović M., Radovanović M., Dai W. Dynamic Time Warping: Itakura vs Sakoe-Chiba, IEEE International Symposium On Innovations in Intelligent SysTems and Applications (INISTA), (2019). https://ieeexplore.ieee.org/document/8778300
4. Haridas A.V., Marimuthu R., Sivakumar V.G. A critical review and analysis on techniques of speech recognition: the road ahead, Int. J. Knowl. Base. Intell. Eng. Syst., V.22, (2018), pp.39-57.
5. Itakura F. Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, V.23, N.1, (1975), pp.67-72.
6. Jiang Sh., Chen Z. Application of dynamic time warping optimization algorithm in speech recognition of machine translation. RESEARCH ARTICLE, V.9, N.11, https://doi.org/10.1016/j.heliyon.2023.e21625
7. Keogh E.J., Pazzani M.J. Derivative Dynamic Time Warping. Proceedings of the 2001 SIAM International Conference on Data Mining, https://doi.org/10.1137/1.9781611972719.1
8. Kerimov A.B. Accuracy comparison of signal recognition methods on the example of a family of successively horizontally displaced curves, Informatics and Control Problems, V.42, N.2, (2022), pp.80-91.
9. Kerimov A.B. Comparison of some signal recognition methods for their adequacy, Proceedings of the 8th International Conference on Control and Optimization with Industrial Applications, V.I, 24-26 August, Baku, Azerbaijan, (2022).
10. Linh L.H., Hai N.T., Thuyen N.V., Mai T.T., Toi V.V. MFCC-DTW Algorithm for Speech Recognition in an Intelligent Wheelchair. 5th International Conference on Biomedical Engineering in Vietnam, (2014), pp. 417-421.
11. Novozhilov B.M. Calculation of the derivative of an analog signal in a programmable logic controller, Aerospace scientific journal of Moscow State Technical University. N.E. Bauman, Electron. Magazine, N.4, (2016), pp.1-12. (In Russian).
12. Rajeev R., Abhishek T. Analysis of feature extraction techniques for speech recognition system, Int. J. Innovative Technol. Explor. Eng., V.8, (2019), pp.197-200.
13. Rzayev R.R., Kerimov A.B. Signal recognition using weighted additive convolution of evaluation criteria. The Springer Series “Lecture Notes in Networks and Systems”, 758, V.2, (2023), pp.407-416, https://doi.org/10.1007/978-3-031-39777-6_49
14. Sakoe H., Chiba S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, V.ASSP-26, N.1, (1978), pp.43-49.
15. Zhi-Qiang U., Jia-Qi Z., Xin W., Zi-Wei L., Yong L. Improved algorithm of DTW in speech recognition. IOP Conference Series: Materials Science and Engineering, 1311, 011001, V.563, N.5, (2024), DOI 10.1088/1757-899X/563/5/052072.