УДК 395.521

В.В. Киселев, А.Г. Давыдов, А.В. Ткаченя

АЛГОРИТМ СРАВНЕНИЯ ФОНОГРАММ В ОБУЧАЮЩИХ СИСТЕМАХ

Рассматривается алгоритм определения сходства между фонограммами на основе линейных спектральных частот и метода динамического программирования. Решается задача сравнения фонограмм как задача различения анализируемого сигнала, в котором присутствуют искажения и помехи с неизвестным началом, длительностью и нелинейным масштабом времени, и образца произношения. Приводится функциональная блок-схема рассматриваемого алгоритма.

Введение

Речевые технологии предлагают пользователям широкий спектр автоматизированных услуг, одной из которых является автоматическая оценка степени сходства фонограмм.

Задача сравнения фонограмм широко востребована в современном мире, особенно в сфере здравоохранения [1] для определения патологических изменений в речевом тракте, в образовании [2] для контроля правильности произношения слов и выражений при обучении языкам, а также может быть использована для оценки качества канала передачи речевых данных [3].

В статье рассматривается метод определения сходства между фонограммами на основе метода динамического программирования (DTW – от англ. dynamic time warping). Суть метода заключается в последовательном сравнении анализируемой записи с образцом. При помощи метода динамического программирования происходит сравнение массивов линейных спектральных частот (LSF – от англ. linear spectral frequency) анализируемой записи и образца произношения. Данный подход часто используется для построения простых систем распознавания речи [4, 5].

1. Алгоритм сравнения фонограмм

Анализ фонограмм выполняется в соответствии со схемой, согласно которой анализируемая запись сравнивается с каждым из образцов правильного произношения, а конечный результат анализа вычисляется как медианное значение результатов сравнения отдельных фонограмм (рис. 1). Выбор медианного значения в качестве результата анализа требуется для получения устойчивой оценки степени сходства фонограмм и обусловлен необходимостью исключения чрезмерной адаптации к конкретному образцу произношения.

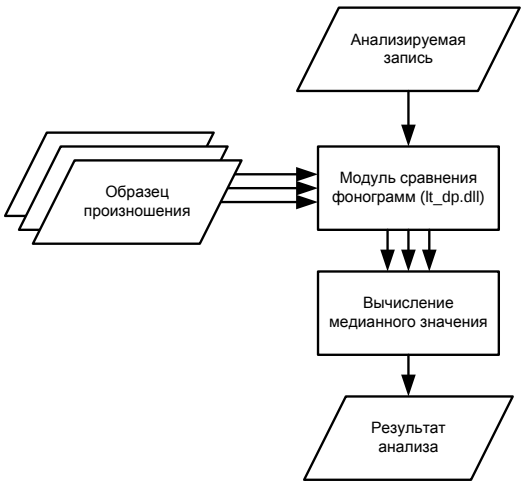

Рис. 1. Блок-схема алгоритма сравнения фонограмм

Сравнение каждой фонограммы образца произношения с анализируемой записью выполняется в соответствии со схемой, приведенной на рис. 2.

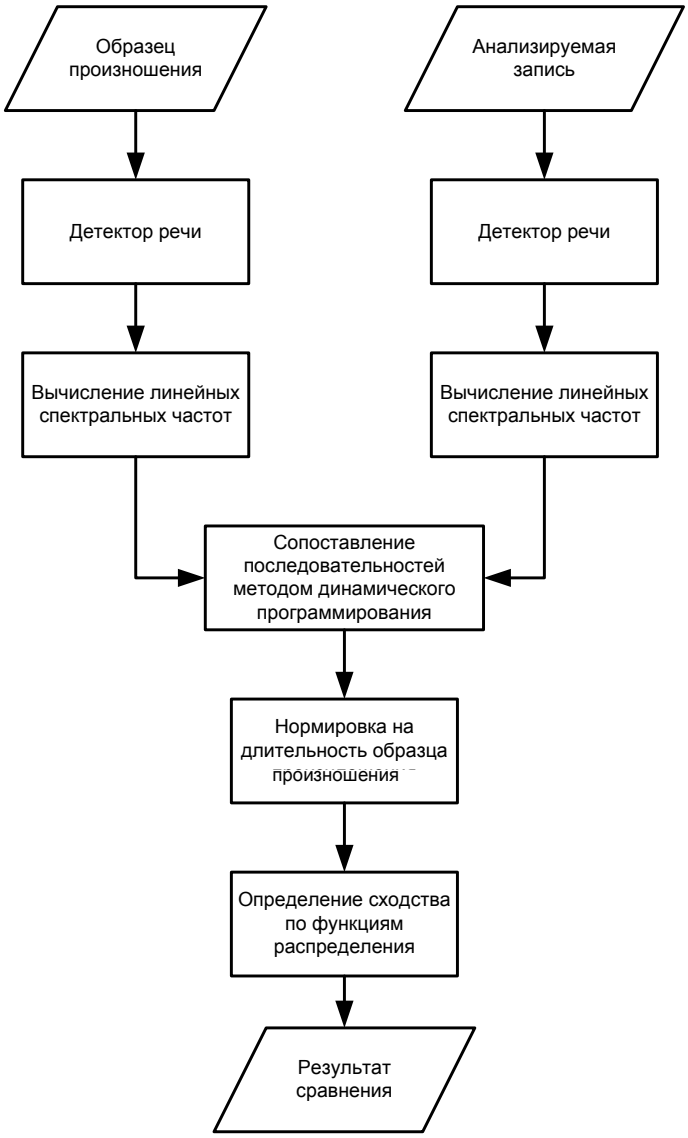

Рис. 2. Блок-схема сравнения двух фонограмм

Особенность предложенного алгоритма сравнения двух фонограмм заключается в использовании блока нормирования на длительность образца произношения, что позволяет снизить временные и аппаратные затраты на сравнение анализируемой записи с образцом произношения. Учитывая необходимость наличия не менее 20 образцов произношения для получения корректной оценки, введение нормирования позволяет использовать данный алгоритм в интернет-сервисах, требующих быстрого анализа большого количества запросов пользователей.

2. Метод детектирования речи

Сравнение двух фонограмм начинается с определения границ речевых участков в анализируемой записи и образце произношения при помощи детектора речи (рис. 2).

Метод детектирования речи [6] основан на анализе оценки мощности сигнала в полосе от 300 до 4000 Гц и требует выполнения следующих правил:

1. Уровнем шума считается минимальное значение оценки мощности сигнала за последние 20 с.

2. Уровнем речи считается максимальное значение оценки мощности сигнала за последние 20 с, но не меньшее уровня шума 20 дБ.

3. За порог детектирования речи принимается среднее между уровнем шума и уровнем речи.

4. Длительность пауз между речевыми участками не может быть менее 50 мс (все паузы длительностью менее 50 мс считаются речью).

5. Длительность речевых участков не может быть менее 70 мс (все речевые участки длительностью менее 70 мс считаются шумом).

На основе данных правил определяются участки начала и окончания речи в сравниваемых записях.

Полоса анализируемого сигнала и остальные параметры были получены в ходе анализа корпуса русской речи, в котором предварительно была произведена экспертная разметка на участки «тишина – речь». Для этого весь корпус был разбит на обучающую и тестируемую выборки в соотношении 80 на 20 % соответственно. Тестирование алгоритма детектирования речи с вышеописанными параметрами дало 98 % совпадения автоматической разметки с экспертной.

Для повышения устойчивости работы блока сопоставления методом динамического программирования было предложено в начале и в конце анализируемых фонограмм оставлять блоки шума длительностью 200 мс.

3. Выбор информативных признаков

На детектированных участках сигнала вычисляются значения линейных спектральных частот (LSF-коэффициенты) [7]. Выбор LSF-коэффициентов в качестве информативных признаков обусловлен их низкой чувствительностью к шумам и сравнительно невысокой вычислительной сложностью. Это позволяет анализировать зашумленные данные, записанные при помощи компьютерной гарнитуры, и не накладывает жестких ограничений на производительность при работе в реальном времени.

Чтобы получить LSF-коэффициенты, p корней полинома предсказания $A_p(z)$ отражаются на единичную окружность посредством двух z-преобразований $(p+1)$-го порядка:

$$P_{p+1}(z) = A_p(z) + z^{-(p+1)} A_p\left(z^{-1}\right), \quad Q_{p+1}(z) = A_p(z) - z^{-(p+1)} A_p\left(z^{-1}\right),$$

т. е.
$$A_p(z) = \frac{P_{p+1}(z) + Q_{p+1}(z)}{2}.$$

LSF-коэффициенты представляют собой угловые позиции корней $P(z)$ и $Q(z)$ на единичной окружности в диапазоне $0 \le \omega_i \le \pi$.

4. Сравнение фонограмм

Сопоставление последовательностей линейных спектральных частот осуществляется методом динамического программирования [8]. DTW позволяет найти оптимальное соответствие между двумя заданными последовательностями. При этом мера подобия таких последовательностей не зависит от изменения нелинейного масштаба времени. Эти свойства DWT наилучшим образом подходят для решения поставленной задачи сравнения фонограмм.

С целью формирования матрицы локальных расстояний d_{ij} для каждой пары сравниваемых LSF-коэффициентов вычисляется L1-метрика:

$$d_{ij} = \sum_{n=1}^{p} \left| LSF_{in} - LSF_{jn} \right|.$$

Определение матрицы интегральных расстояний D_{ij} выполняется с использованием локальных ограничений Итакуры [9]:

$$D_{ij} = \min\begin{Bmatrix} D_{i-2,j-1} + d_{i-1,j} \\ D_{i-1,j-1} \\ D_{i-1,j-2} + d_{i,j-1} \end{Bmatrix} + d_{ij}.$$

Расстоянием между сравниваемыми записями является значение матрицы интегральных расстояний с максимальными индексами $D_{\max\_i,\max\_j}$.

5. Нормировка интегрального расстояния

Нормировка интегрального расстояния на длительность анализируемой записи позволяет в первом приближении использовать функции распределения, полученные для других фонограмм, и таким образом избежать трудоемкой процедуры определения фактических функций распределения интегральных расстояний:

$$D_n = D_{\max\_i,\max\_j} / N.$$

Определение значения сходства Sim между фонограммами выполняется на основе определения значений функций распределения «своих» (правильное произношение фонограммы – F_{fr}, сплошная линия), «чужих» (неправильное произношение – F_{foe}, пунктир) и их точек пересечения $(q_{ee}; F_{ee})$ (рис. 3):

$$\mathrm{Sim} = \begin{cases} \dfrac{1 + \left(F_{ee} - F_{fr}\right)/F_{ee}}{2}, \text{ если } D_n \le q_{ee}; \\ \dfrac{1 - \left(F_{foe} - F_{ee}\right)/\left(1 - F_{ee}\right)}{2}, \text{ если } q_{ee} < D_n. \end{cases}$$

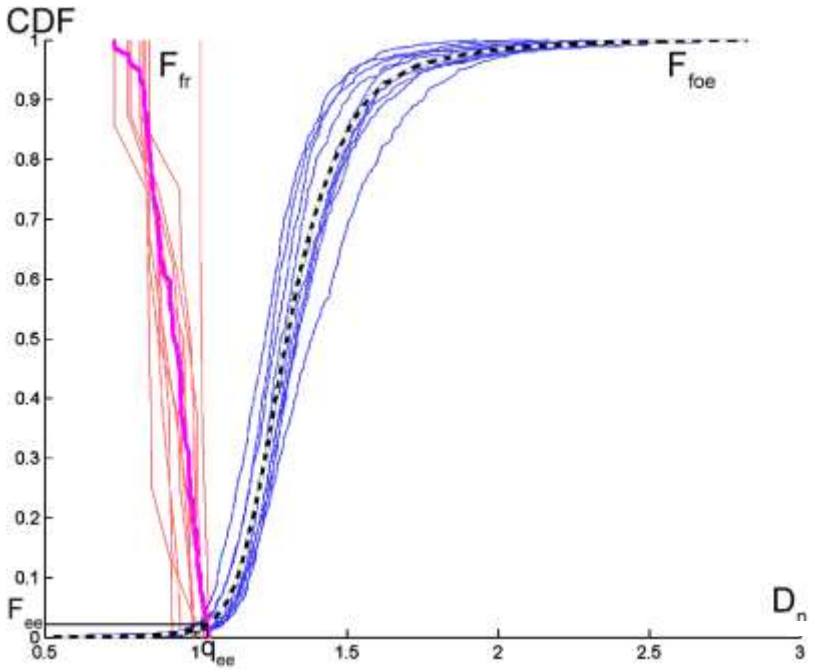

Рис. 3. Функции распределения «свои» и «чужие» для фразы «акклиматизироваться в Константинополе»

6. Проверка эффективности разработанного алгоритма сравнения фонограмм

Разработанный алгоритм сравнения фонограмм предназначен для контроля правильности произношения слов и выражений при обучении языкам. Работа алгоритма предусматривает запись пользователем требуемой речевой фонограммы и получение комплексной оценки меры подобия записанного сигнала с заданными образцами произношения (см. рис. 1).

Для корректной работы алгоритма сравнения фонограмм необходимо 20 образцов правильного произношения для каждой из фонограмм. Для этого был записан речевой корпус, состоящий из 10 дикторов (пять мужчин и пять женщин), каждый из которых произносит заданную фонограмму четыре раза: два раза правильно – образец произношения («свои») и два раза с ошибками – вариант произношения фонограммы, который считается неправильным («чужие»). На основе этих данных получаются функции распределения, которые впоследствии используются для определения сходства между записанным сигналом и образцами произношения (см. рис. 2).

С целью проверки эффективности разработанного алгоритма в рамках эксперимента было проведено его сравнение с работой аналогичного алгоритма фирмы «ИстраСофт» из программного продукта «Профессор Хиггинс. Русский без акцента!» [10], в котором реализована автоматическая оценка правильности произношения одного слова (таблица). Были выбраны три типа фонограмм: одиночное слово, фраза (до семи слов) и скороговорка. В тестировании принимали участие четыре диктора (двое мужчин и две женщины), не вошедших в обучающую выборку.

Проверка эффективности работы алгоритма сравнения фонограмм проводилась на файлах, записанных при наличии соотношения сигнал/шум 15 и 30 дБ (SNR), клиппированного сигнала (clipping), одиночной ошибки (1 miss) и множественной ошибки (*N* miss).

Степень сходства анализируемых записей при различных шумах и искажениях, %

Программный комплекс	SNR 15 dB	SNR 30 dB	clipping	1 miss	*N* miss
Одно слово					
Алгоритм сравнения фонограмм	56	91	47	75	43
Профессор Хиггинс. Русский без акцента!	63	89	57	71	54
Фраза (до семи слов)					
Алгоритм сравнения фонограмм	54	87	36	81	45
Скороговорка					
Алгоритм сравнения фонограмм	51	90	37	84	49

Заключение

На основе изложенной в статье теории был разработан алгоритм сравнения фонограмм, который предназначен для оценки правильности произношения звуков, слов и фраз. Разработанный алгоритм характеризуется низкими временными и аппаратными затратами, что позволяет применять его в системах, обслуживающих одновременно большое количество пользователей. По сравнению с уже имеющимися продуктами («Профессор Хиггинс. Русский без акцента!») он позволяет получать автоматическую оценку правильности произношения фонограмм, состоящих из нескольких слов.

Результаты эксперимента показали, что разработанный алгоритм не уступает в эффективности сравнения одного слова алгоритмам, существующим на данный момент. Вместе с тем клиппированность сигнала сильно снижает эффективность работы алгоритма. Наличие большого количества шумов в сигнале приводит к падению правильности сравнения фонограмм на 40 %. Для одиночной ошибки результат сильно зависит от общей длины фонограммы: чем длиннее фраза, тем сильнее нивелируется одиночная ошибка.

К недостаткам разработанного алгоритма сравнения фонограмм можно отнести трудоемкий и долгий этап обучения системы.

Список литературы

1. Stanley, T. Improving L1-specific phonological error diagnosis in computer assisted pronunciation training / T. Stanley, K. Hacioglu // Proc. of Interspeech. – Portland, Oregon, 2012. – P. 153–159.

2. Rose, R. Verifying session level pronunciation accuracy in a speech therapy application / R. Rose, S.-C. Yin, Y. Tang // Proc. of Interspeech. – Portland, Oregon, 2012. – P. 267–272.

3. Perceived speech quality estimation using DTW algorithm / I. Kraljevski [et al.] // Telfor Journal. – 2009. – № 1. – P. 25–31.

4. Performance of DTW speech recognizer on packet switched network / I. Kraljevski [et al.] // Proc. of 7th ETAI Conf. – Ohrid, Macedonia, 2005. – P. 89–96.

5. Paliwal, K.K. On the use of line spectral frequency parameters for speech recognition / K.K. Paliwal // Proc. of Digital. Signal Processing 2. – Bombay, India, 1992. – P. 80–87.

6. Sakhnov, K. Approach for energy-based voice detector with adaptive scaling factor / K. Sakhnov, E. Verteletskaya, B. Simak // IAENG Intern. Journal of Computer Science. – 2009. – № 36 (4). – P. 48–53.

7. Kabal, P. The computation of line spectral frequencies using chebyshev polynomials / P. Kabal, R.P. Ramachandran // IEEE Trans. Acoustics, Speech, Signal Processing. – 1986. – № 34 (6). – P. 1419–1426.

8. Rabiner, L. Fundamentals of speech recognition / L. Rabiner, B.-H. Juang. – NJ, USA, 1993. – 496 p.

9. Keogh, E. Exact indexing of dynamic time warping / E. Keogh, C.A. Ratanamahatana. – USA : University of California–Riverside, 2004. – 417 p.

10. Профессор Хиггинс. Русский без акцента! // ИстраСофт [Электронный ресурс]. – Режим доступа : http://www.istrasoft.ru/ru/programmy/professor-higgins-russkij-bez-akcenta.html. – Дата доступа : 11.09.2013.

Поступила: 12.08.2013

ООО «Речевые технологии»,
Минск, пер. Уральский, 15
e-mail: kiselev-v@speetech.by;
davydov-a@speetech.by;
tkachenia-a@speetech.by

V.V. Kiselev, A.G. Davydau, A.V. Tkachenia

ALGORITHM FOR PHONOGRAM COMPARISON IN E-LEARNING SYSTEMS

An algorithm for determination of the degree of phonogram matching on the basis of linear spectral frequency and dynamic time warping is considered. The problem of phonogram comparison is handled as the problem of recognizing the analyzed speech signal, which contains a lot of distortion and noise with unknown beginning and duration, non-linear timescale, and unknown pronunciation sample. A functional diagram of the algorithm is presented.