Алгоритм сравнения фонограмм на основе каналонезависимых информативных признаков

Киселёв В.В.,

Ткаченя А.В.

Проводится исследование информативных признаков с целью формирования каналонезависимого пространства признаков для повышения эффективности голосового анализа, для решения задачи определения сходства между фонограммами на основе метода динамического программирования.

• *голосовой анализ* • *машинное обучение* • *выбор информативных признаков* • *мел-кепстральные коэффициенты* • *метод динамического программирования*

The research of informative feature vectors to form channel-independent feature space to improve the efficiency of speech analysis for solving the problem of comparing phonograms on the basis of dynamic time warping.

• *speech analysis* • *machine learning* • *feature selection* • *Mel-frequency cepstral coefficients* • *Dynamic Time Warping*

Введение

Важнейший этап при создании систем автоматического голосового анализа — выделение оптимального набора информативных признаков, так как их выбор оказывает значительное влияние на эффективность классификации. При решении большинства прикладных задач анализу подвергаются голосовые данные, полученные при различных условиях записи. Изменение характеристик канала приводит к изменению анализируемого пространства признаков, что ведёт к снижению эффективности классификации.

Для снижения влияния характеристик канала на эффективность работы систем голосового анализа необходимо использовать каналонезависимые информативные признаки. В последнее время исследования в этом направлении приобрели особую актуальность [1, 2]. Тем не менее, большинство существующих способов получения каналонезависимых информативных признаков характеризуются большими времен-

ными и аппаратными затратами. Это затрудняет их использование в задачах, требующих анализа сигнала в реальном времени.

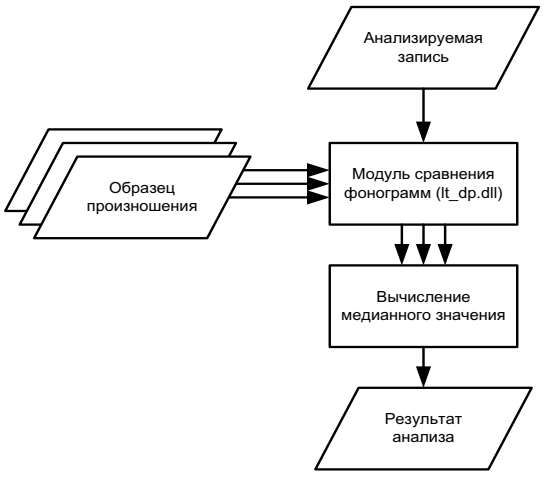

Рис. 1. Блок-схема алгоритма сравнения фонограмм

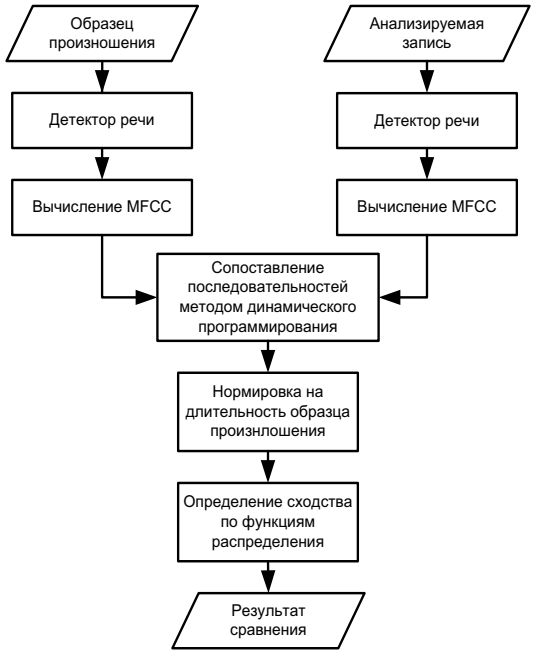

Рис. 2. Блок-схема сравнения двух фонограмм

В экспериментальной части данной работы приводится сравнение полученной эффективности для случая использования исходных информативных признаков и полученных каналонезависимых на примере задачи определения сходства между фонограммами на основе метода динамического программирования (Dynamic Time Warping — DTW). Суть метода заключается в последовательном сравнении анализируемой записи с образцом. При помощи метода динамического программирования происходит сравнение массивов информативных признаков анализируемой записи и образца произношения. Данный подход часто используется для построения простых систем распознавания речи [3, 4].

1. Алгоритм сравнения фонограмм

Анализ фонограмм выполняется в соответствии со схемой, приведённой на *рис. 1.*

Так, анализируемая запись сравнивается с каждым из образцов правильного произношения, а конечный результат анализа вычисляется как медианное значение результатов сравнений отдельных фонограмм. Выбор медианного значения в качестве результата анализа требуется для получения устойчивой оценки степени сходства фонограмм и обусловлен необходимостью исключения чрезмерной адаптации на конкретный образец произношения.

Порядок сравнения массивов мел-кепстральных коэффициентов (Mel-frequency cepstral coefficients — MFCC) каждой фонограммы-образца произношения с анализируемой записью схематично показан на *рис. 2.*

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.

2. Выбор информативных признаков

Восприятие тонов человеческим ухом носит нелинейный характер. Количественное описание этой нелинейности задаётся так называемой мел-шкалой, определяемой эмпирическим соотношением:

$$B(f) = 1125\ln(1 + f/700)$$

Речевой сигнал можно представить как свёртку двух функций изначального вида акустической волны $s(t)$ (исходного сигнала) и фильтра $h(t)$ (зависящего от параметров голосового тракта), параметры которого должны быть оценены как

$$f(t) = s(t) \otimes h(t)$$

В частотной области получаем

$$F(\omega) = S(\omega) \cdot H(\omega)$$

Полученный спектр $F(\omega)$ нужно расположить на мел-шкале *(рис. 3)*:

$$W_m(\omega) = \begin{cases} 0 & \omega < f[m-1] \\ \dfrac{2(\omega - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & f[m-1] \le \omega \le f[m] \\ \dfrac{2(f[m+1] - \omega)}{(f[m+1] - f[m-1])(f[m+1] - f[m])} & f[m] \le \omega \le f[m+1] \\ 0 & \omega > f[m+1] \end{cases}$$

где m =1,2,...,M; M — количество треугольных фильтров.

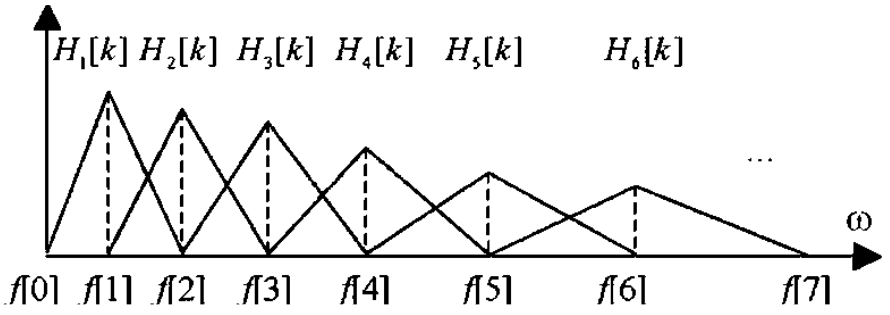

Рис. 3. Треугольные фильтры, используемые для получения мел-спектра

Для дискретного случая $f[m]$ может быть вычислено следующим образом:

$$f[m] = \left(\frac{N}{F_S}\right) B^{-1}\left(B(f_1) + m\frac{B(f_h) - B(f_1)}{M+1}\right), \quad B^{-1}(b) = 700\left(\exp(b/1125) - 1\right),$$

где f_l и f_h — минимальная и максимальная частоты фильтров в Гц; F_S — частота дискретизации сигнала в Гц; M — количество фильтров; N — размер БПФ.

Найдём логарифмированную энергию сигнала на выходе каждого из фильтров:

$$X[m] = \ln\left[\int_{-\infty}^{\infty} |F(\omega)|^2 W_m(\omega) d\omega\right], \qquad 0 \le m < M$$

Для фильтров с плавной передаточной функцией MFCC является гомоморфным преобразованием, что позволяет получить выражение

$$X[m] = \int_{-\infty}^{\infty} \ln\left(|F(\omega)|^2 W_m(\omega)\right) d\omega, \qquad 0 \le m < M$$

Тогда спектр $F(\omega)$ можно представить как сумму исходного сигнала и фильтра:

$$\ln\left(|F(\omega)|^2\right) = \ln\left(S^2(\omega) \cdot H^2(\omega)\right) = \ln S^2(\omega) + \ln H^2(\omega)$$

Теперь необходимо преобразовать эту сумму так, чтобы получить непересекающиеся наборы характеристик исходного сигнала и фильтра. Для этого вводится преобразование кепстров:

$$C[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X[m] \left| e^{i\omega n} d\omega \right. \text{ — вещественный кепстр;}$$

$$C[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X[m]\, e^{i\omega n} d\omega \text{ — комплексный кепстр,}$$

где ω — частота в радианах.

Такой подход позволяет получить характеристики речевого сигнала (мел-кепстральные коэффициенты — MFCC), которые минимально зависят от индивидуальных особенностей говорящего, а значит, могут быть очень полезны в задачах распознавания [5].

3. Построение каналонезависимых информативных признаков

При решении прикладных задач анализируются данные, полученные при различных условиях записи, что ведёт к изменению анализируемого пространства признаков и, в свою очередь, к снижению эффективности классификации. Для достижения устойчивой работы и снижения разброса эффективности голосового анализа необходимо использовать каналонезависимые информативные признаки.

Так как в сигнале могут присутствовать шумы, то вначале каждая запись проходит детектор речи, основанный на анализе оценки мощности сигнала в полосе от 300 до 4000 Гц [6], с целью выделения речевых (участки сигнала, в которых присутствует речь) и неречевых участков (участки с шумом). Далее на полученных речевых участках осуществляется классификация вокализованных и невокализованных участков, основанная на нахождении нормализованной кросс-корреляционной функции [7].

Часто в литературе можно встретить подход к нормировке параметров канала связи (адаптации коэффициентов наблюдений) посредством вычитания средних значений коэффициентов вещественного кепстра. Такой подход позволяет эффективно бороться с мультипликативными искажениями, вносимыми различными каналами связи.

Вычитание средних значений мел-кепстральных коэффициентов вместо вычитания средних значений коэффициентов вещественного кепстра накладывает определённые ограничения на виды допустимых мультипликативных искажений, однако является более эффективным в вычислительном плане. При этом встречаются различные способы оценки среднего значения мел-кепстральных коэффициентов:

1. *Оценка средних значений на неречевых участках.* Этот способ позволяет эффективно бороться с мультипликативными искажениями канала связи, сохраняя информацию об индивидуальных голосовых характеристиках диктора.

2. *Оценка средних значений как на вокализованных, так и на невокализованных участках речи.*

3. *Оценка средних значений только на вокализованных участках речи.* Позволяет нормировать коэффициенты наблюдений как к каналу связи, так и к голосу диктора. При этом за счёт того, что средние значения оцениваются только на вокализованных участках речи, дисперсии оценок оказываются меньше, чем при оценке средних на вокализованных и невокализованных участках речи.

При необходимости работы в реальном времени популярным способом вычитания среднего является применение фильтра с коэффициентами $b = [1; -1]$, $a = [1; -0{,}97]$. При этом инициализация фильтра выполняется таким образом, чтобы $x_0 = x_1$, $y_0 = 0$.

Амплитудно-частотная характеристика (АЧХ) и фазо-частотная характеристика (ФЧХ) такого фильтра приведены на *рис. 4*.

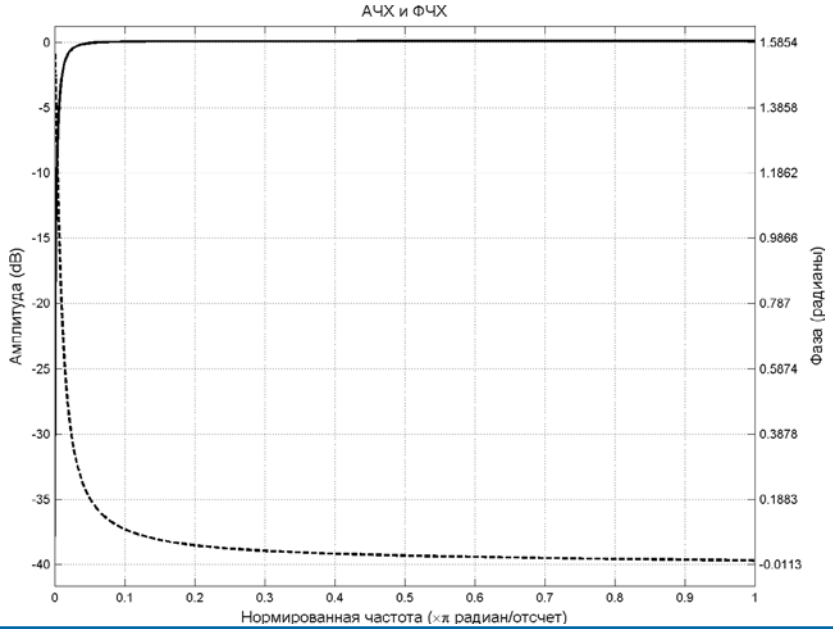

Рис. 4. АЧХ (сплошная линия) и ФЧХ (пунктирная линия) фильтра

Для того чтобы информативные признаки стали каналонезависимыми, было предложено провести оценку средних значений только на вокализованных участках речи. Это было обусловлено тем, что вычитание средних значений коэффициентов вещественного кепстра приводит к возрастанию вычислительных затрат, что затрудняет использование этого метода в реальном времени. При оценке средних значений на неречевых участках сохраняется информация об индивидуальных голосовых характеристиках дик-

тора, что снижает эффективность сравнения фонограмм. А случай оценки средних значений на вокализованных и невокализованных участках речи даёт большую дисперсию оценок, что также приводит к снижению эффективности.

Такой шаг позволяет эффективно бороться с мультипликативными искажениями, вносимыми различными каналами связи. Полученные каналонезависимые мел-кепстральные коэффициенты нормированы как к каналу связи, так и к голосу диктора, что значительно повышает эффективность алгоритма сравнения фонограмм.

4. Сравнение фонограмм

Сопоставление последовательностей мел-кепстральных коэффициентов осуществляется методом динамического программирования [5]. DTW позволяет найти оптимальное соответствие между двумя заданными последовательностями. При этом мера подобия этих 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 MFCC-коэффициентов вычисляется L1-метрика:

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

Определение матрицы интегральных расстояний D_{ij} выполняется с использованием локальных ограничений Итакуры [8].

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

Расстоянием между сравниваемыми записями является значение матрицы интегральных расстояний с максимальными индексами D_{max_i,max_j}.

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

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

Определение значения сходства S_{im} 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. 5):*

$$Sim = \begin{cases} \dfrac{1 + (F_e - F_f)/F_e}{2}, & \text{если} \quad D_n \le q_e; \\ \dfrac{1 - (F_{foe} - F_e)/(1 - F_e)}{2}, & \text{если} \quad q_e < D_n. \end{cases}$$

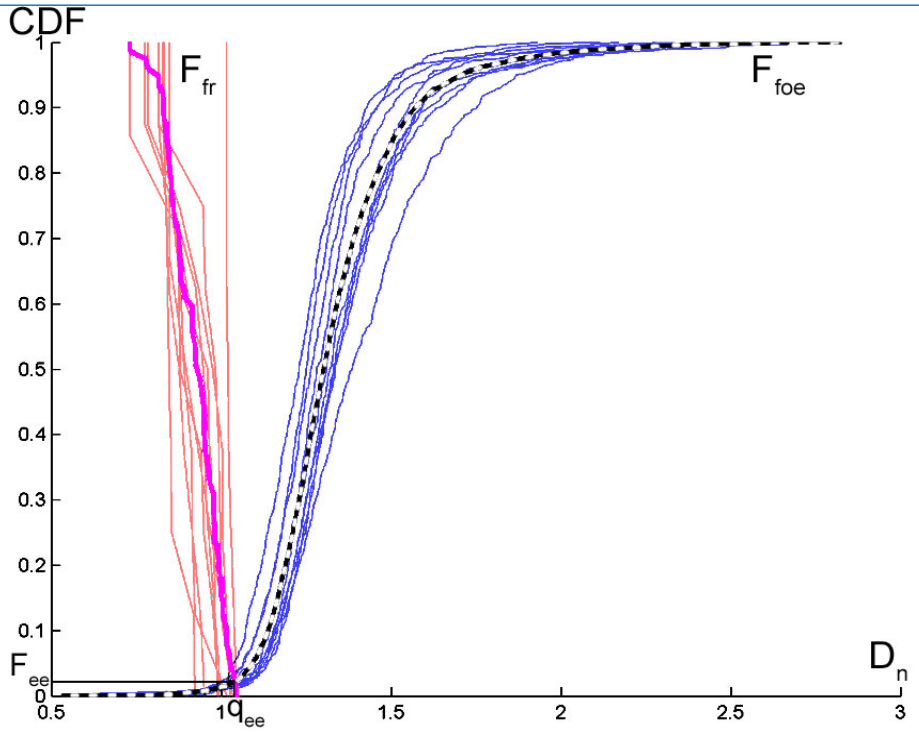

Рис. 5. Функции распределения «своих» и «чужих» для фразы «акклиматизироваться в Константинополе»

5. Результаты эксперимента

Рассмотрим результаты экспериментального исследования описанного способа формирования каналонезависимого пространства признаков для построения классификатора и сравнения эффективности предложенного алгоритма оценки сходства фонограмм при использовании исходных и каналонезависимых информативных признаков (таблица).

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)*.

База образцов произношения записывается на конденсаторном микрофоне BEHRINGER C-2 (с частотным диапазоном 20-20000 Гц и соотношением сигнал/шум 75 дБ) с использованием внешней звуковой карты Creative E-MU 0202 USB 2.0. Тестирование алгоритма сравнения фонограмм осуществлялось на гарнитуре A4Tech HS-5P (с частотным диапазоном 20-20000 Гц и соотношением сигнал/шум 97 дБ), подключённой к встроенной звуковой карте.

Для проведения эксперимента 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 вошедшие в обучающую выборку.

Проверка эффективности работы алгоритма оценки сходства фонограмм проводилась на файлах, записанных при соотношении сигнал/шум 15 и 30 дБ (SNR), клиппированном сигнале (clipping), одиночной ошибке произнесения (1 miss), множественной ошибке произнесения (N miss).

Заключение

Предложенный способ построения каналонезависимых информативных признаков характеризуется низкими временными и аппаратными затратами. Это позволяет их использовать в системах голосового анализа без значительного снижения производительности конечного программного комплекса.

Использование каналонезависимых информативных признаков приводит к повышению точности разделения правильного и неправильного произношения фонограммы (см. таблицу). При этом эффективность классификации зашумленных и клиппированных сигналов значительно возросла в среднем на 20–25%.

Таблица

Степень сходства анализируемых записей при различных шумах и искажениях

Информативный признак	SNR 15 dB	SNR 30 dB	clipping	1 miss	N miss
Одно слово					
MFCC-коэффициенты	57%	92%	46%	75%	42%
Каналонезависимые MFCC-коэффициенты	79%	93%	68%	77%	44%
Фраза (до семи слов)					
MFCC-коэффициенты	54%	88%	37%	80%	45%
Каналонезависимые MFCC-коэффициенты	76%	90%	60%	79%	40%
Скороговорка					
MFCC-коэффициенты	53%	89%	38%	83%	49%
Каналонезависимые MFCC-коэффициенты	74%	91%	63%	80%	42%

В качестве дальнейшей работы представляется целесообразным протестировать эффективность применения описанных каналонезависимых информативных признаков для классификации психоэмоционального состояния человека по его речи.

Литература

1. Amplitude Modulation Filters as Feature Sets for Robust ASR: Constant Absolute or Relative Bandwidth? / Moritz N. [et al.] // Interspeech 2012. Portland, Oregon, 2012. September 9–13. P. 76–83.
2. Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition / Meyer Bernd T. [et al.] // Interspeech 2012. Portland, Oregon, 2012. September 9–13.P. 132–141.
3. Performance of DTW speech recognizer on packet switched network / I. Kraljevski [et al.] // Proc. of 7th ETAI Conf. Ohrid, Macedonia, 2005. P. 89–96.
4. Paliwal K.K. On the use of line spectral frequency parameters for speech recognition // Proc. of Digital. SignalProcessing2. Bombay, India, 1992. P. 80–87.
5. Rabiner L., Juang B.-H. Fundamentals of speech recognition. Prentice-Hall, Inc.Upper Saddle River, NJ, USA,1993.
6. Sakhnov K., Verteletskaya E., Simak B. Approach for energy-based voice detector with adaptive scaling factor // IAENG Intern. Journal of Computer Science. 2009. № 36 (4). P. 48–53.
7. Talkin D. A Robust Algorithm for Pitch Tracking // Speech Coding and Synthesis. 1995. P. 495–518.
8. Keogh E., Ratanamahatana C.A. Exact indexing of dynamic time warping. USA: University of California — Riverside, 2004.

Сведения об авторах

Киселёв Виталий Владимирович,

директор ООО «Речевые технологии», кандидат технических наук, г. Минск, Беларусь. С 1999 г. профессионально занимается системами синтеза и распознавания речи, диалоговыми речевыми системами. Автор более 25 научных публикаций в области речевых технологий. Основные научные интересы связаны с системами обработкой и анализом текста и речи, системами синтеза, распознавания речи, поиска ключевых слов.
E-mail: kiselev-v@speetech.by

Ткаченя Андрей Владимирович,

младший научный сотрудник ООО «Речевые технологии». Область научных интересов — системы анализа и индексирования аудиосигналов, скрытые Марковские модели в задачах распознавания речи. Беларусь, г. Минск, пер. Уральский, 15.
E-mail: tkachenia-a@speetech.by