УДК 519.86

В. В. Киселёв, А. В. Ткаченя, М. В. Хитров

РАЗРАБОТКА КАНАЛОНЕЗАВИСИМЫХ ИНФОРМАТИВНЫХ ПРИЗНАКОВ

Исследованы информативные признаки речи с целью формирования каналонезависимого пространства признаков для повышения эффективности функционирования системы распознавания дикторов. Экспериментально определен оптимальный набор каналонезависимых информативных признаков для решения задачи выявления сходства между фонограммами на основе метода динамического программирования.

Ключевые слова: *голосовой анализ, машинное обучение, выбор информативных признаков, мел-частотные кепстральные коэффициенты, метод динамического программирования.*

Введение. Важнейшим этапом в создании систем автоматического голосового анализа является выделение оптимального набора информативных признаков. При решении большинства прикладных задач анализу подвергаются голосовые данные диктора, полученные при различных условиях записи. Изменение характеристик канала приводит к изменению анализируемого пространства признаков, что снижает эффективность классификации дикторов.

Цель предлагаемой работы — снижение влияния характеристик канала на эффективность работы систем голосового анализа. Для достижения цели необходимо использовать каналонезависимые информативные признаки. В последнее время исследования в этом направлении приобрели особую актуальность [1—3]. Тем не менее, большинство существующих способов получения каналонезависимых информативных признаков характеризуются большими временными и аппаратными затратами, что затрудняет их использование в задачах, требующих анализа сигнала в реальном масштабе времени.

В настоящей работе сравнивается эффективность для случая использования исходных информативных и полученных каналонезависимых признаков на примере задачи выявления сходства между фонограммами. Для этого применяется метод динамического программирования (DTW), заключающийся в последовательном сравнении анализируемой записи с образцом. При помощи DTW происходит сравнение массивов информативных признаков анализируемой записи и образца произношения. Данный подход часто используется при построении простых систем распознавания речи [4, 5].

Алгоритм сравнения фонограмм. Анализ фонограмм выполняется в соответствии с блок-схемой, приведенной на рис. 1.

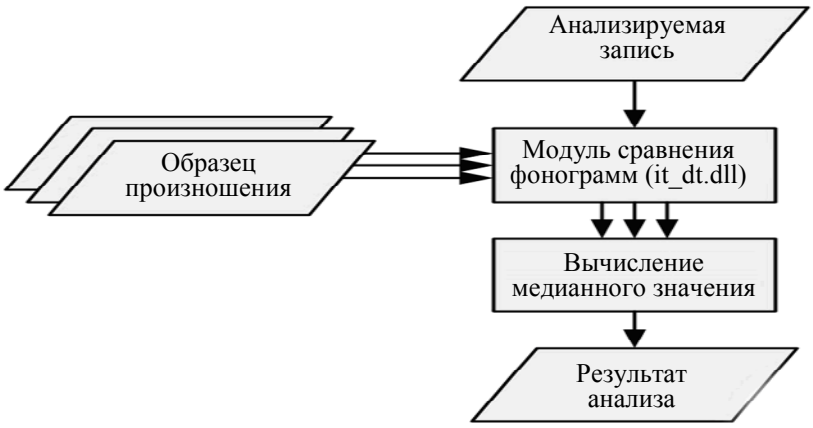

Рис. 1

Из рисунка видно, что анализируемая запись сравнивается с каждым из образцов правильного произношения, а конечный результат анализа вычисляется как медианное значение результатов сравнения отдельных фонограмм. Использование медианного значения позволяет получить устойчивую оценку степени сходства фонограмм и обусловлено необходимостью исключения чрезмерной адаптации к конкретному образцу произношения.

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

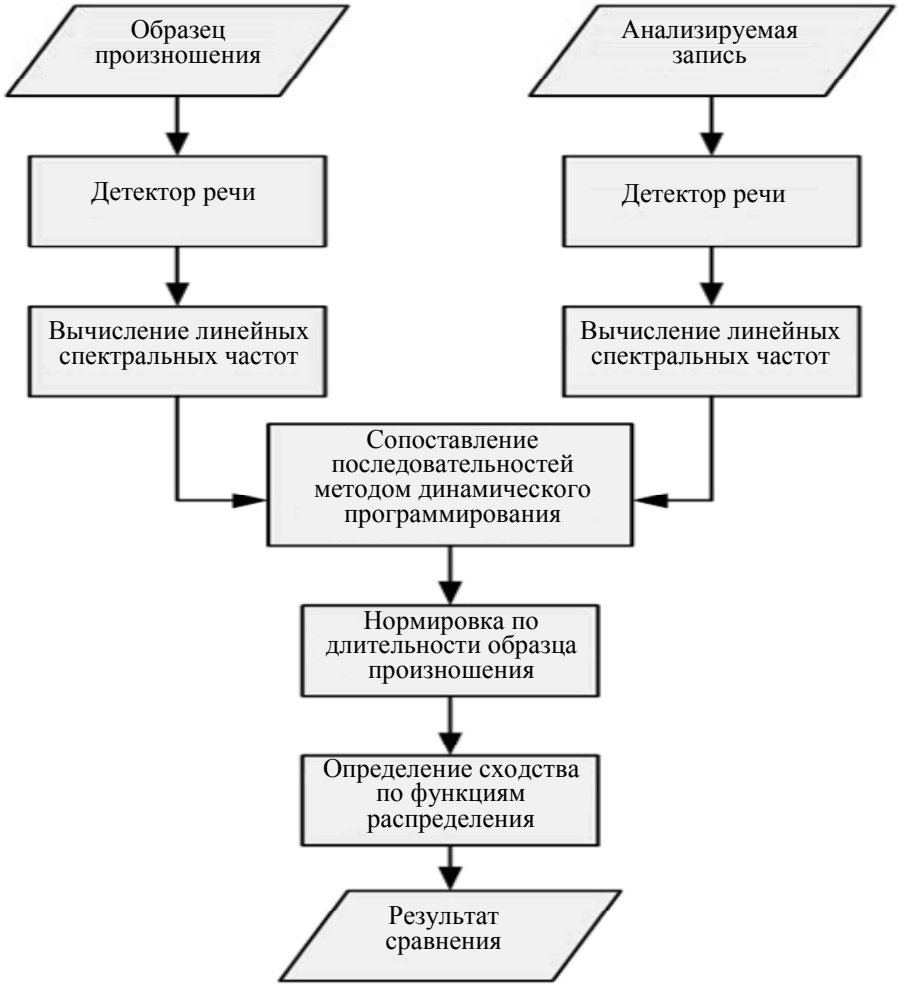

Рис. 2

Особенность предложенного алгоритма сравнения двух фонограмм заключается в использовании блока нормирования по длительности образца произношения, что позволяет снизить временные и аппаратные затраты на сопоставление анализируемой записи с образцом.

Выбор информативных признаков. Известно, что чувствительность человека к звуковому сигналу зависит от частоты сигнала: чем ниже частота, тем чувствительность выше. В 1937 г. была выведена формула, по которой можно перевести частоту (f) в герцах в частоту в мелах (m):

$$m = 1127,01048\ln(1+f/700), \quad f = 700(e^{m/1127,01048}-1).$$

Сигнал представляется как свертка двух функций: исходного сигнала и фильтра, параметры которого должны быть оценены. Необходимо разделить эти отдельные компоненты при помощи преобразования

$$x * h = \hat{x} + \hat{h}\,.$$

Для этого вводится кепстральное преобразование
— вещественный кепстральный коэффициент:

$$C[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|X(e^{i\omega})\right| e^{i\omega n} d\omega\,;$$

— комплексный кепстральный коэффициент:

$$C[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(X(e^{i\omega})) e^{i\omega n} d\omega\,,$$

где $X(e^{i\omega})$ — спектр сигнала; ω — частота (в радианах).

Такой подход позволяет получить характеристики речевого сигнала (мел-частотные кепстральные коэффициенты, MFCC), которые минимально зависят от индивидуальных особенностей говорящего, а значит, могут быть очень полезны в задачах распознавания [6].

Так как при решении прикладных задач анализируются данные, полученные в различных условиях записи, изменяется анализируемое пространство признаков и снижается эффективность классификации. Для достижения робастности голосового анализа в системах распознавания диктора необходимо использовать каналонезависимые информативные признаки.

Часто в литературе нормировка параметров канала связи (адаптация коэффициентов наблюдений) выполняется посредством вычитания средних значений коэффициентов вещественного кепстра. Такой подход позволяет эффективно бороться с мультипликативными искажениями, вносимыми различными каналами связи.

Вычитание средних значений MFCC вместо вычитания средних значений коэффициентов вещественного кепстра накладывает определенные ограничения на виды допустимых мультипликативных искажений, однако более эффективно в вычислительном плане.

Возможны 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, что позволяет нормировать коэффициенты наблюдений как к каналу связи, так и к голосу диктора. За счет 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 режиме реального времени для вычитания среднего часто применяется фильтр с коэффициентами **b** = [1 –1], **a**=[1 –0,97]. При этом инициализация фильтра выполняется таким образом, чтобы x_0=x_1, y_0=0. АЧХ (*2*) и ФЧХ (*1*) такого фильтра приведены на рис. 3 ($\overline{f} = f\pi$ радиан/отсчет).

Для того чтобы информативные признаки стали каналонезависимыми, было предложено провести оценку средних значений только на вокализованных участках речи. Такой шаг позволяет вышеописанные мел-частотные кепстральные коэффициенты, сильно зависящие от

характеристик канала, сделать каналонезависимыми и значительно повысить эффективность использующих их систем.

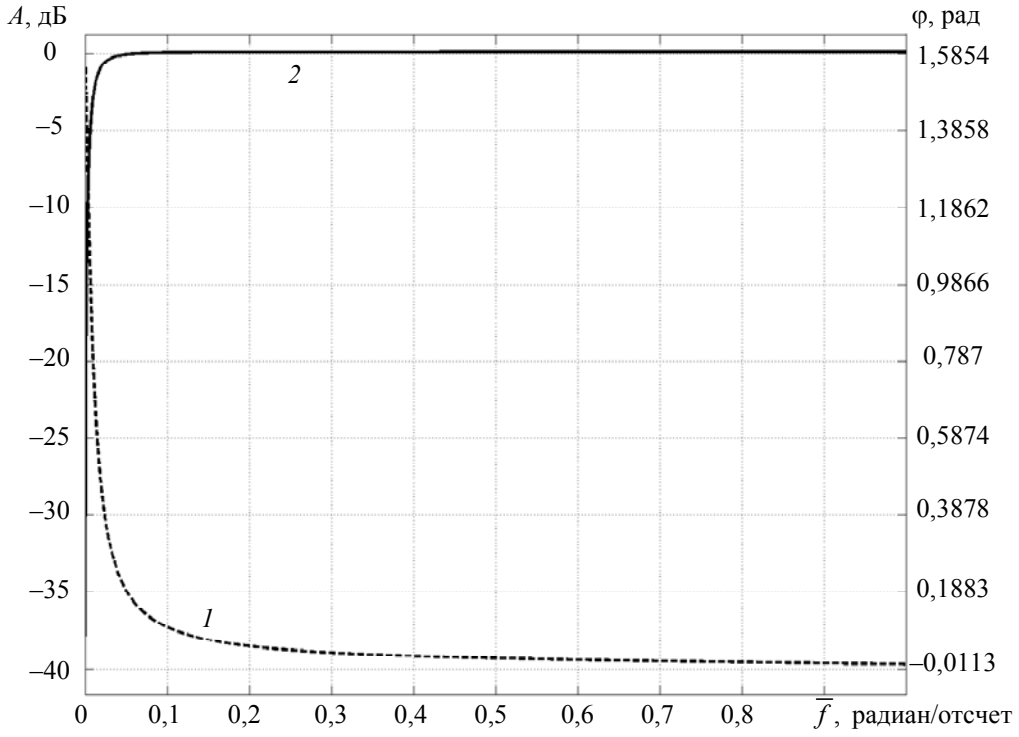

Рис. 3

Сравнение фонограмм. Ранее последовательности MFCC сопоставлялись методом динамического программирования [6]. DTW позволяет найти максимальное подобие двух заданных последовательностей, при этом мера их подобия 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 вычисляется L1-метрика:

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

Определение матрицы интегральных расстояний D_{ij} выполняется с использованием локальных ограничений Итакуры [7]:

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

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

Результаты экспериментов. 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 проведения эксперимента были выбраны три типа фонограмм: одиночное слово, фраза (до 7 слов) и скороговорка. В тестировании принимали участие 4 диктора (2 мужчины и 2 женщины), не вошедшие в обучающую выборку. Проверка эффективности работы алгоритма оценки сходства фонограмм проводилась на файлах, записанных при следующих условиях: соотношение сигнал/шум (SNR) 15 и 30 дБ, клиппирование сигнала (clipping) [8], одиночная ошибка (1 miss), множественная ошибка (N miss). Результаты тестирования приведены в таблице.

Степень сходства анализируемых записей при различных шумах и искажениях

Информативный признак	SNR 15 дБ	SNR 30 дБ	Clipping	1 miss	N miss
1 слово					
MFCC	57	92	46	75	42
Каналонезависимые MFCC	79	93	68	77	44
Фраза					
MFCC	54	88	37	80	45
Каналонезависимые MFCC	76	90	60	79	40
Скороговорка					
MFCC	53	89	38	83	49
Каналонезависимые MFCC	74	91	63	80	42

Заключение. В статье предложен метод формирования каналонезависимого пространства признаков классификатора на основе MFCC. Было проведено экспериментальное исследование эффективности предложенного метода, включающее определение оптимального набора параметров и построение классификатора для выявления сходства фонограмм. Такой способ построения каналонезависимых информативных признаков характеризуется низкими временными и аппаратными затратами, что позволяет их использовать в системах голосового анализа без значительного снижения производительности конечного программного комплекса.

Как видно из таблицы, использование каналонезависимых информативных признаков приводит к повышению точности разделения правильного и неправильного произношения фонограммы. При этом эффективность классификации зашумленных и клиппированных сигналов значительно возросла: в среднем на 20—25 %.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Moritz N., Anemüller J., Kollmeier B.* Amplitude Modulation Filters as Feature Sets for Robust ASR: Constant Absolute or Relative Bandwidth? // Proc. 13th Annual Conf. of the Intern. Speech Communication Association (Interspeech-2012). Portland, Oregon, USA, 2012. P. 1230—1233.

2. *Meyer B. T., Spille C., Kollmeier B., Morgan N.* Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition // Proc. 13th Annual Conference of the International Speech Communication Association (Interspeech-2012). Portland, Oregon, USA, 2012. P. 1258—1261.

3. *Матвеев Ю. Н.* Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 47—51.

4. *Kraljevski I., Gacovski Z., Arsenovski S., Mihajlov M.* Performance of DTW Speech Recognizer on Packet Switched Network // Proc. VII ETAI Conf. Ohrid, Macedonia, 2005. P. 16—20.

5. *Paliwal K. K.* On the Use of line Spectral Frequency Parameters for Speech Recognition // Digital Signal Processing. 1992. Vol. 2. P. 80—87.

6. *Rabiner L., Biing-Hwang Juang*. Fundamentals of speech recognition. Inc. Upper Saddle River, NJ, USA: Prentice-Hall, 1993. 496 p.

7. *Keogh E., Ratanamahatana C.A.* Exact indexing of dynamic time warping // Knowledge and Information Systems. 2005. Vol. 7, Is. 3. P. 358—386.

8. *Алейник С. В., Матвеев Ю. Н., Раев А. Н.* Метод оценки уровня клиппирования речевого сигнала // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 3 (79). С. 79—83.

Сведения об авторах

Виталий Владимирович Киселёв — ООО „Речевые технологии“, Минск; директор; E-mail: kiselev-v@speechpro.com

Андрей Владимирович Ткаченя — ООО „Речевые технологии“, Минск; младший научный сотрудник; E-mail: tkachenia-a@speechpro.com

Михаил Васильевич Хитров — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; генеральный директор; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; заведующий кафедрой; E-mail: khitrov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 004.934

Н. А. Томашенко, Ю. Ю. Хохлов

ИССЛЕДОВАНИЕ ПРОБЛЕМЫ СБАЛАНСИРОВАННОСТИ ДАННЫХ ПРИ ПОСТРОЕНИИ АКУСТИЧЕСКИХ МОДЕЛЕЙ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Исследована проблема сбалансированности данных при обучении акустических моделей для автоматического распознавания речи. Предложена метрика, позволяющая при кластеризации состояний трифонов явно учитывать влияние количества данных в кластере. Экспериментально доказано, что использование такого подхода позволяет повысить качество распознавания речи.

Ключевые слова: *автоматическое распознавание речи, GMM-HMM, обучение акустических моделей, связывание состояний, сбалансированность данных, кластеризация, трифоны.*

Введение. Качество системы автоматического распознавания речи в значительной степени определяется характеристиками используемых в ней акустических моделей. В настоящее время в области распознавания речи обычно применяются статистические подходы, при этом свойства акустических моделей во многом зависят от характеристик речевой базы данных, на которой эти модели были обучены. Одна из наиболее распространенных проблем, связанных с речевыми базами данных, — различие объемов (несбалансированность) данных, приходящихся на разные акустические классы, что может оказывать серьезное влияние на классифицирующую способность моделей [1]. В частности, отсутствие необходимого количества данных в обучающей выборке для определенных моделей усложняет получение надежной оценки параметров этих моделей.

Проблеме несбалансированности классов уделено много внимания в литературе по машинному обучению (см., например, [2]). Несмотря на то что многие алгоритмы обучения предполагают сбалансированность данных, это условие не всегда выполняется для реальных приложений, когда одни классы представлены большим количеством данных в обучающей выборке, а другие — всего несколькими элементами. Этой особенностью отличаются и речевые базы данных, используемые при построении акустических моделей.