УДК 681.3.06

Анализ эффективности метода нечёткого сопоставления образов для распознавания изолированных слов

Бондаренко И.Ю., Федяев О.И., к.т.н., доц.
Донецкий национальный технический университет
bond005@yandex.ru, fedyaev@r5.dgtu.donetsk.ua

Рассматривается применение метода нечёткого сопоставления образов к решению задачи автоматического распознавания изолированных слов. Для решения проблемы временной нестационарности сравниваемых образов предлагается алгоритм нелинейного выравнивания длин сравниваемых образов на основе метода DTW. Проводится сравнительный анализ эффективности различных модификаций метода нечёткого сопоставления образов при линейном и нелинейном выравнивании длин сопоставляемых образов.

Введение

Речевой интерфейс, как более естественный для человека, приобретает всё большую востребованность в современных человеко-машинных системах. Об этом свидетельствует и возросшее число коммерческих разработок систем, использующих речевой интерфейс. Так, NaturallySpeaking фирмы Dragon System позволяет редактировать и форматировать текст с помощью собственного текстового процессора без использования клавиатуры и мыши. Компания IBM разработала аналогичную программу, позволяющую осуществлять речевой ввод и форматирование текста в текстовом процессоре MS Word. На практике эти программы показывают недостаточно высокие результаты (при тестировании точность не достигла даже 90% [1]). Корпорация Microsoft также начала активно заниматься внедрением речевого интерфейса в свои программные продукты. С помощью компонент MS Speech API программист может организовывать речевой интерфейс в любой прикладной программе. Но, несмотря на заявленную в фирменной документации точность распознавания 95% [2], на практике точность распознавания и, следовательно, надёжность программных систем, использующих речевой интерфейс на основе MS Speech API, невысока.

В основе построения речевого интерфейса лежит задача распознавания речи, для решения которой, несмотря на множество предложенных способов, не найден приемлемый метод. Наметились основные направления, базирующиеся на вероятностном, метрическом и нейросетевом подходах. Также перспективен для решения трудноформализуемых задач, к которым относится задача распознавания речи, подход на основе нечёткой логики [3]. В работе [4] описан метод нечёткого сопоставления образов и приведена высокая оценка его эффективности в распознавании английских, немецких и японских слов. Однако изменения длительности одинаковых речевых образов, обусловленные различной скоростью произношения звуков одних и тех же слов, в указанной работе рассматриваются только как линейные в целях уменьшения объёма вычислений. Но изменение длин речевых образов в общем случае является нелинейным [5], поэтому целесообразно оценить учёт этой нелинейности при построении модели временной нормализации.

В данной работе предлагается повысить эффективность метода нечёткого сопоставления образов с помощью использования алгоритма DTW [6] для нелинейной временной нормализации сравниваемых образов. Алгоритм DTW, основанный на принципах динамического программирования, устанавливает временное соответствие между звуками сопоставляемых речевых образов. Проводится сравнительный анализ эффективности работы системы распознавания, использующей линейную нормализацию, и системы, выравнивающей длины образов по алгоритму DTW.

Получение информативных признаков речевого сигнала

Речевой сигнал представляется в виде двумерного спектрального временного образа (СВО), получаемого с помощью оконного преобразования Фурье (рис.1а). Такой образ отражает изменение по времени амплитуд заданных частотных составляющих речевого сигнала и хорошо выражает особенности речи, что даёт возможность его использовать для автоматического распознавания произносимых слов [4]. СВО позволяет выделить местоположение резонансных частот, т.е. локальных выбросов, что является определяющей особенностью речевого сигнала [4]. На этом основании СВО можно преобразовать к двоичному виду, не теряя указанных информативных признаков речи, с помощью следующей замены: 1 – на месте локального выброса, 0 – в других местах. Полученный образ называют двоичным спектральным

временным образом (ДСВО) и используют его как отражение особенностей речевого сигнала (рис. 1б).

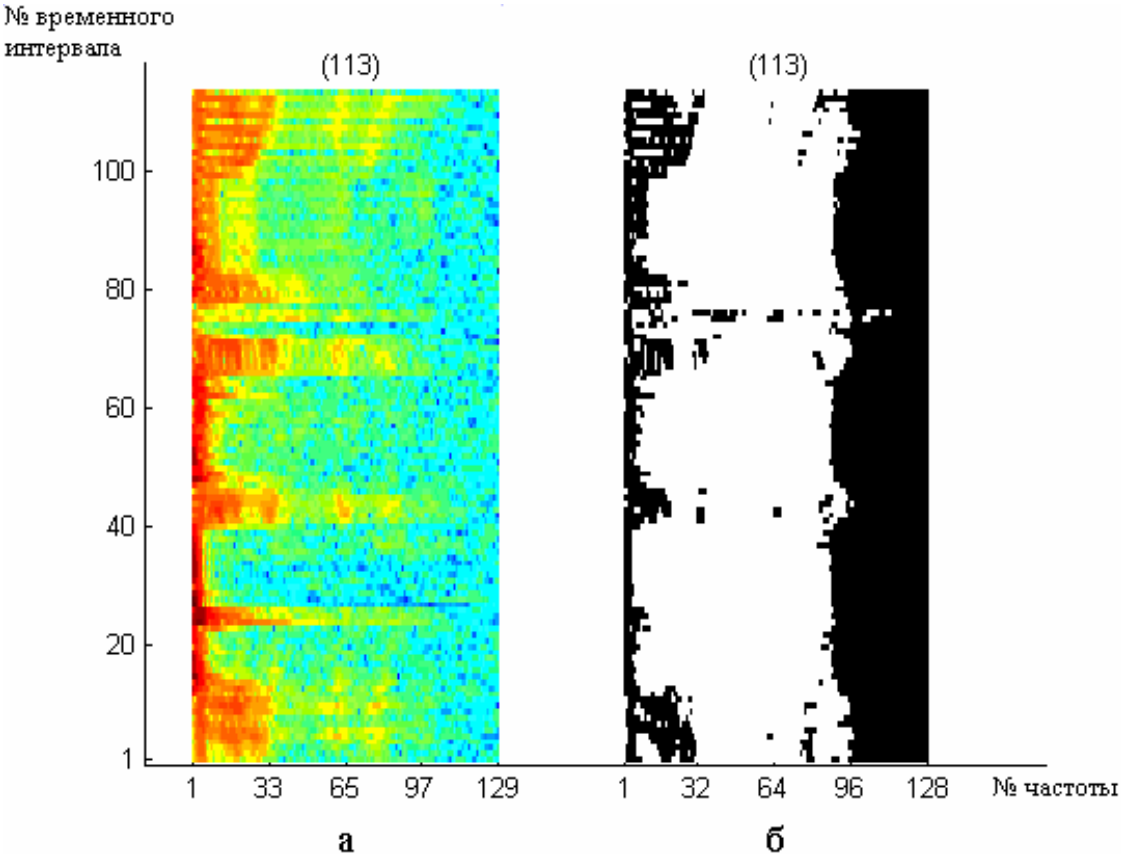

Рисунок 1 – Пример спектрально-временного представления слова «автоформат»: а – СВО; б – ДСВО

В качестве единиц речи рассматриваются слова, набор которых определяет словарный состав речевого общения.

Временное выравнивание речевых образов

Различные реализации речевых образов, относящихся к одному и тому же классу, могут значительно отличаться друг от друга по длительности (рис.2). Это связано с нестабильностью темпа речи диктора, вызванной влиянием интонации, акцента и т.п. Для корректного сопоставления речевых образов необходимо производить их выравнивание по длине. Выравнивание путём линейного сжатия или растяжения одной реализации слова до величины другой решает задачу лишь частично, так как не учитывается одно важное свойство речевого сигнала – неравномерность его протекания во времени [5]. Это свойство

речи выражается в неравномерном изменении длительности звуков слова при изменении длительности слова в целом. Поэтому сопоставление целесообразно выполнять с помощью нелинейной временной нормализации.

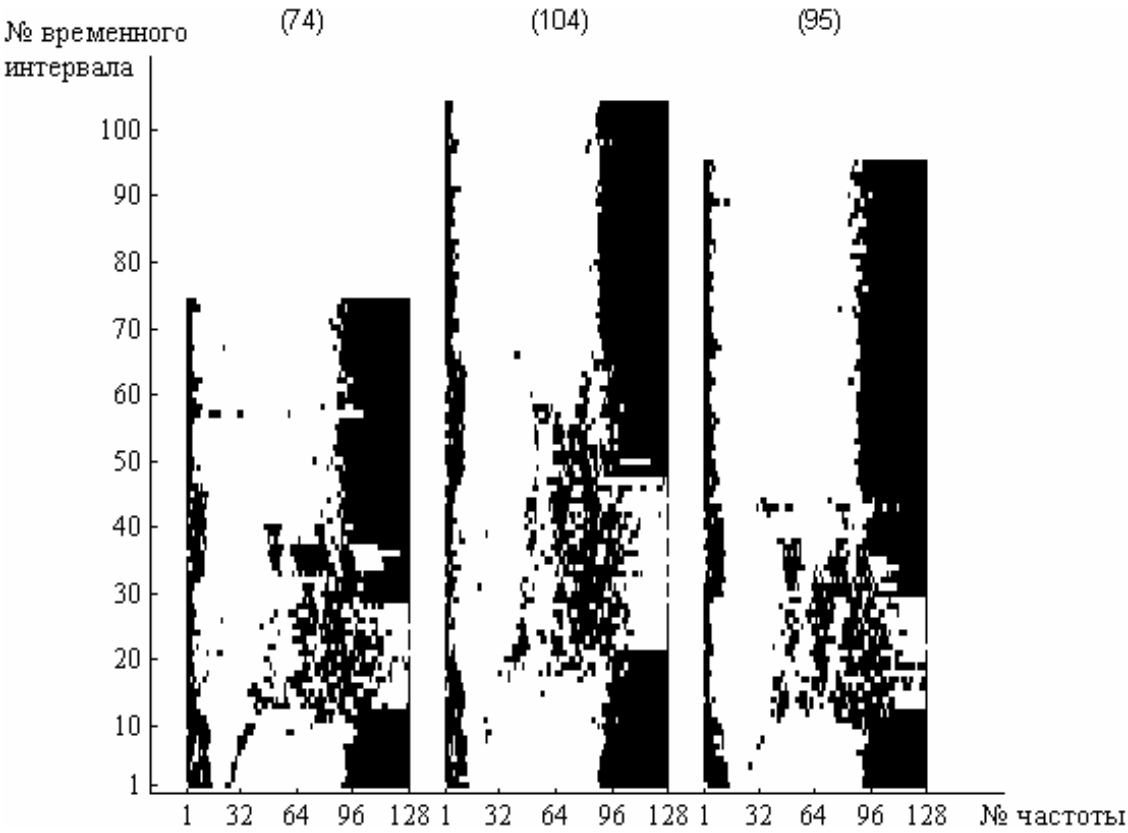

Рисунок 2 – ДСВО различных реализаций слова «Курсив»

Для нелинейного выравнивания сопоставляемых образов использовался алгоритм, основанный на определении наилучшего соответствия входных и эталонных речевых образов, известный как метод DTW [6]. Суть алгоритма заключается в следующем. Обозначим евклидово расстояние между i-й строкой матрицы входного ДСВО и j-й строкой матрицы эталона как D_{ij}. Для нахождения строк матрицы входного ДСВО, наилучшим образом соответствующих строкам матрицы эталона, определялась матрица C размера (M*N) по следующим формулам:

$C(1, 1) = D_{11}$;
$C(\mathrm{i}, 1) = D_{i1} + C(i - 1, 1)$, i=2..M;
$C(1, j) = D_{1j} + C(1, j - 1)$, j=2..N;
$C(i, j) = D_{ij} + \min[C(i - 1, j), C(i - 1, j - 1), C(i, j - 1)]$, i=2..M, j=2..N ,

где M – количество строк матрицы входного ДСВО; N – количество строк матрицы эталона.

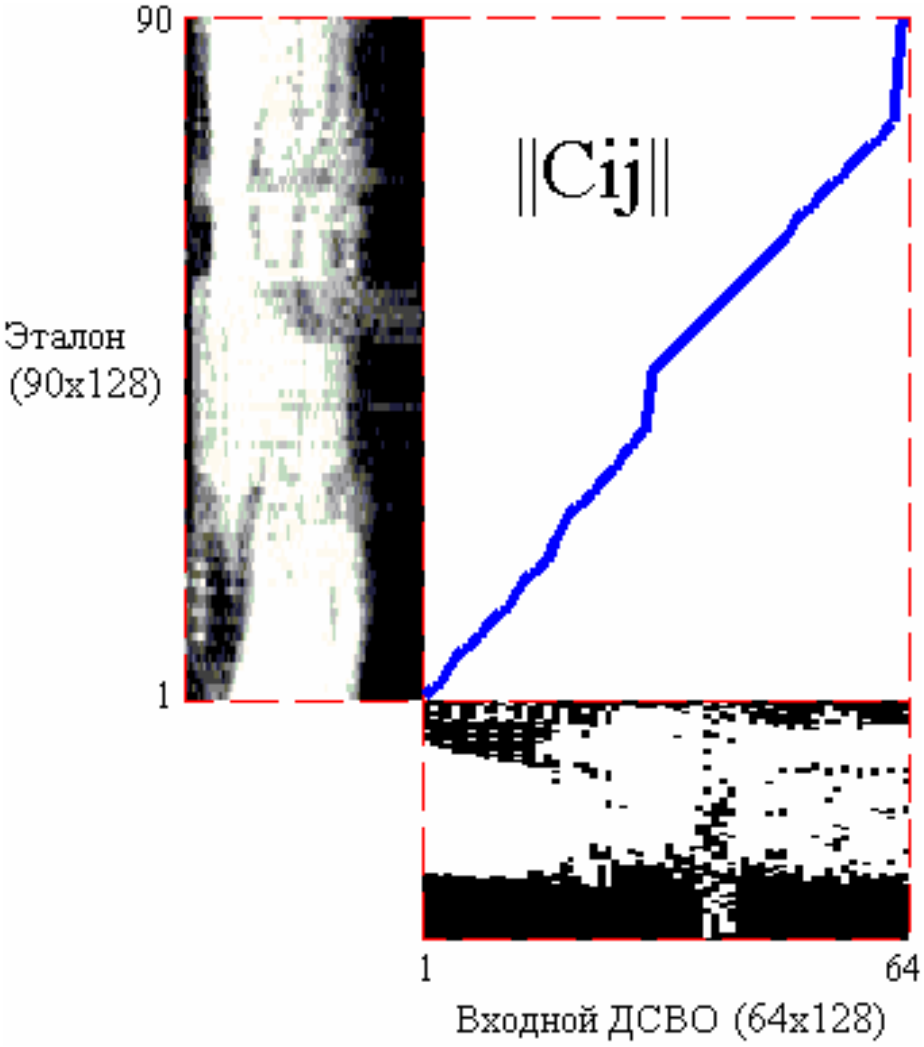

Рисунок 3 – Графическое отображение процесса выравнивания входного ДСВО и эталона по алгоритму DTW

На рис. 3 ломаной линией соединены элементы матрицы C, отвечающие наиболее соответствующим строкам входного ДСВО и эталона слова «Маркеры». Вертикальному отрезку соответствует случай, когда несколько строк матрицы эталона соответствуют одной строке матрицы входного ДСВО. Горизонтальному отрезку соответствует случай, когда несколько строк матрицы ДСВО соответствуют одной строке матрицы эталона. Таким образом, в отличие от алгоритма линейного приведения длин, данный алгоритм

обеспечивает выравнивание только спектрально подобных фрагментов входного ДСВО и эталонного образа.

Метод нечёткого сопоставления речевых образов

Для распознавания изолированных слов, нормализованных по времени, применялся метод нечёткого сопоставления с эталоном [4]. Эталонные образы для каждого слова словаря формировались как среднее арифметическое ДСВО различных вариантов произношения данного слова. В результате формируется бинарное нечёткое отношение между множеством F (номеров частот f) и множеством T (номеров временных интервалов t) в виде:

$$f \in F, t \in T : F \; \boldsymbol{R} \; T,$$

где R – нечёткое отношение, которое ставит каждой паре элементов $(f,t) \in F \times T$ величину функции принадлежности $\mu_R(x,y) \in [0,1]$.

Обозначим число записанных слов через n, множество слов через $I = \{i_1, i_2, ... i_n\}$ и множество нечётких отношений, характерных для каждого слова, через $R = \{r_1, r_2, ... r_n\}$. Входной неизвестный образ y рассматривается как обычное (чёткое) отношение между множеством частот и множеством временных интервалов. Для него вычисляются степени сходства S_j с каждым нечётким отношением r_j. Результатом распознавания является слово j, такое, что

$$j = \max_{j \in I} \{S_j\}.$$

Степень подобия вычисляется по следующей формуле:

$$S_j = \frac{D_j}{\overline{D_j}},$$

где

$$D_j = \int r(f,t) \wedge y(f,t) df dt, \quad \overline{D_j} = \int \overline{r(f,t)} \wedge y(f,t) df dt.$$

В дискретном случае имеет место

$$D_j = \sum_t \sum_f r(f,t) \wedge y(f,t), \; \overline{D_j} = \sum_t \sum_f \overline{r(f,t)} \wedge y(f,t)$$

Влияние методов выравнивания на качество распознавания

Были проведены экспериментальные исследования, направленные на определение качества распознавания слов русской речи по методу нечёткого сопоставления при линейном и нелинейном выравнивании образов. Для эксперимента использовалась речевая однодикторная база данных, включавшая в себя звукозаписи 6 речевых команд управления текстовым процессором: «Автоформат», «Жирный», «Курсив», «Маркеры», «Найти», «Нумерация». Каждая речевая команда была представлена 30 реализациями, 15 из которых использовались для обучения системы, а 15 – для тестирования. Таким образом, мощности обучающего и тестового множеств составили 90 различных реализаций вышеперечисленных 6 слов.

Результаты распознавания слов тестового множества по методу нечёткого сопоставления с использованием линейного временного выравнивания представлены в табл. 1, а с использованием временного выравнивания по алгоритму DTW – в табл. 2.

Таблица 1. **Результаты тестирования системы с линейным выравниванием**

	Автоформат	Жирный	Курсив	Маркеры	Найти	Нумерация	**Итого, %**
Автоформат	15	0	0	0	0	0	**100,00**
Жирный	0	14	0	0	1	0	**93,33**
Курсив	0	0	15	0	0	0	**100,00**
Маркеры	0	0	0	13	0	2	**86,67**
Найти	0	0	0	0	15	0	**100,00**
Нумерация	0	0	0	0	0	15	**100,00**
Качество распознавания составило **96,67%**							

Таблица 2. **Результаты тестирования системы с DTW-выравниванием**

	Автоформат	Жирный	Курсив	Маркеры	Найти	Нумерация	**Итого, %**
Автоформат	15	0	0	0	0	0	**100,00**
Жирный	0	15	0	0	0	0	**100,00**
Курсив	0	0	15	0	0	0	**100,00**
Маркеры	0	0	0	15	0	0	**100,00**
Найти	0	0	0	0	15	0	**100,00**
Нумерация	0	0	0	0	0	15	**100,00**
Качество распознавания составило **100,00%**							

Заключение

В результате исследований установлено, что метод нечёткого сопоставления эффективен для распознавания изолированных слов и словосочетаний русского языка, в том числе речевых команд управления программными системами. Сравнительный анализ линейного выравнивания и выравнивания по методу DTW показал, что второй способ за счёт учёта нелинейности временных изменений речевых образов повышает эффективность метода нечёткого сопоставления при распознавании речевых образов. Полученные результаты позволяют использовать разработанные алгоритмы нечёткого сопоставления с выравниванием по методу DTW для создания систем речевого командного управления.

Литература

1. Программы синтеза и распознавания речи. Тестовая лаборатория. – http://art.bdk.com.ru/govor/1listr62t.htm.
2. *Буторин Д.Н*. MS Agent и Speech API в Delphi. – С.-Пб.: BHV. – 2005. – 448 с.
3. *Кофман А*. Введение в теорию нечетких множеств. – М.: Радио и связь. – 1982. – 432 с.
4. *Киедзи Асаи, Дзюндзо Ватада, Сокуке Иваи и др*. Распознавание речи // Прикладные нечёткие системы. Под редакцией Т.Тэрано, К. Асаи, М. Сугено. – М.: «Мир», – 1993. – 157-170 с.
5. *Винцюк Т.К.* Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наукова думка. – 1987. – 264 с.
6. *Stuart N. Wrigley*. Speech Recognition by Dynamic Time Warping. – http://www.dcs.shef.ac.uk/~stu/com326/index.html.