Нечёткое сопоставление образов с оптимальным временным выравниванием для однодикторного и многодикторного распознавания изолированных слов

Федяев О.И., Бондаренко И.Ю.
Кафедра ПМиИ ДонНТУ
fedyaev@r5.dgtu.donetsk.ua

Abstract

O.I.Fedyaev, I.U.Bondarenko, Fuzzy Pattern Matching with Optimal Temporal Alignment for Single-Speaker and Multi-Speaker Isolated Word Recognition. *Fuzzy pattern matching applications are considered for automated isolated word recognition task. To solve a temporal instability problem of speech patterns compared an optimal temporal alignment algorithm is proposed, which is based on maximal similarity criterion. Results of experimental research for single-speaker and multi-speaker recognition are described on a set of 105 Russian words using a proposed algorithm.*

Введение

Использование речевых команд в контуре управления современными техническими системами (в частности, роботами) становится всё более востребованным пользователями таких систем, поскольку устная речь – это наиболее естественное для человека средство обмена информацией, не требующее от него специальной подготовки и доступное людям с нарушениями опорно-двигательного и зрительного аппарата. Применение речевого канала управления в условиях интенсивного обмена информацией между человеком-оператором и управляемым устройством освобождает глаза и руки оператора и позволяет ему сосредоточить всё внимание на ходе процесса управления [1]. Решение задачи распознавания речи для систем речевого управления должно удовлетворять следующим требованиям:

– высокоточное распознавание ограниченного числа изолированных речевых слов – команд управления;

– возможность работы в дикторонезависимом режиме.

Основные направления решения данной задачи базируются на вероятностном, метрическом и нейросетевом подходах. Для решения трудноформализуемых задач, к которым относится и вышеуказанная задача распознавания изолированных речевых слов, также перспективен подход на основе нечёткой логики [2].

Известен ряд способов использования нечёткой логики для распознавания изолированных слов. Так, в работе [3] предложена

гибридная система, основанная на сочетании принципов динамического программирования и нечёткой логики. Используются три типа первичных признаков речевого сигнала: кратковременная энергия, кратковременное число переходов через нуль и кепстр. Векторы признаков нормализуются по времени с помощью алгоритма динамических временных деформаций (dynamic time warping) и подаются на вход системы нечёткого логического вывода, которая вырабатывает решение о принадлежности речевого сигнала одному из слов словаря.

Для распознавания изолированных речевых слов также применяют метод нечёткой векторной квантизации, реализованный в нейросетевом базисе [4]. Здесь речевой сигнал представляется набором векторов кепстральных коэффициентов.

Наиболее перспективный способ, использующий нечёткую логику, основывается на методе нечёткого сопоставления образов [5]. В нём определяющей особенностью речевого сигнала является характер изменения местоположения резонансных частот сигнала во времени. Система распознавания изолированных речевых слов, основанная на этом методе, показала высокий результат распознавания набора японских, немецких и английских слов [5]. Основной проблемой метода нечёткого сопоставления образов, на наш взгляд, является проблема временного выравнивания сопоставляемых образов. Для решения этой проблемы предлагалось применять к сопоставляемым образам алгоритмы линейного временного выравнивания [5] и нелинейного выравнивания, основанного на динамическом программировании [6]. У каждого из этих алгоритмов есть недостатки, влияющие на качество распознавания речевых образов. Так, алгоритм линейного выравнивания не учитывает неравномерность протекания речевого сигнала во времени, а алгоритм нелинейного временного выравнивания требует выполнения длительных вычислений и ведёт к снижению различимости речевых образов, относящихся к различным классам. Поэтому в данной работе для приведения сопоставляемых образов к одинаковой длине предложен новый алгоритм, названный оптимальным временным выравниванием. Критерием оптимальности выравнивания является максимизация степени сходства сопоставляемых образов.

1. Получение информативных признаков речевого сигнала

Речевой сигнал представляется в виде двумерного спектрального временного образа (СВО), получаемого разложением сигнала по частотам на группе 15 полосовых фильтров. Частоты анализа распределены на интервале 200…5000 Гц с шагом 1/3 октавы, добротность полосовых фильтров равна 6. Выходные сигналы фильтров сглаживаются и квантуются выборками по 10 мс. Полученный образ отражает изменение

по времени амплитуд заданных частотных составляющих речевого сигнала и хорошо выражает особенности речи, что даёт возможность использовать его для автоматического распознавания произносимых слов [5].

Известно, что человек произносит слова, изменяя органом речи резонансные частоты, поэтому особенно важной информацией в СВО является местоположение резонансных частот, то есть локальных выбросов [5]. На этом основании СВО можно преобразовать к двоичному виду, сохраняя информативные признаки 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.

Рассмотрим задачу нахождения локальных выбросов в СВО. Пусть $X_p(k)$ − спектр сигнала на p-м фрейме, $Y_p(k)$ − местоположение локальных выбросов в спектре сигнала на p-м фрейме, $k = 1…L$. Определяем номера $B = \{b_i \mid i \in \overline{1,M}\}$ локальных максимумов и номера $S = \{s_j \mid j \in \overline{1,N}\}$ локальных минимумов последовательности $X_p(k)$. Для каждого номера локального максимума b_i определяем x_i – угол между двумя прямыми, первая из которых проведена из точки локального максимума в ближайшую к ней слева точку локального минимума s_j, а вторая – из точки локального максимума в ближайшую к ней справа точку локального минимума s_{j+1} (рис. 1). Считаем, что те из найденных локальных максимумов b_i, для которых значения углов крутизны подъёма x_i не превышают заданный порог α, – это локальные выбросы в спектре сигнала, а углы x_i – углы крутизны этих выбросов.

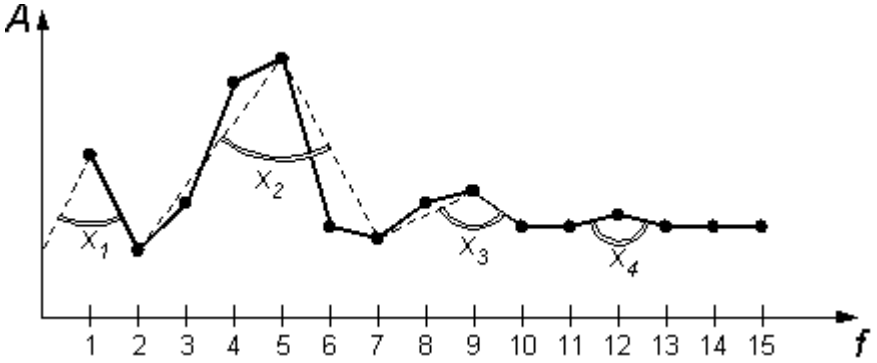

Рисунок 1 – Определение углов x_i крутизны локальных выбросов в спектре

Таким образом, местоположение локальных выбросов в спектре сигнала на p-м фрейме $Y_p(k)$ определяется по следующей формуле:

$$Y_p(k) = \begin{cases} 1, & k \in \{b_i \mid i \in \overline{1,M}\} \;\&\; x_i \le \alpha; \\ 0, & \text{иначе}, \end{cases}$$

где $0 < \alpha \le 180$ (если значения углов крутизны x_i определяются в градусах) − коэффициент, подбираемый экспериментально.

2. Нечёткое сопоставление образов

Распознавание изолированных речевых слов осуществляется на основе метода нечёткого сопоставления образов [5]. ДСВО речевого слова представляет собой бинарное отношение между множеством F (номеров частот f, по которым выполняется спектральный анализ речевого сигнала), и множеством T (номеров временных интервалов t, на которые речевой сигнал квантуется по времени) в виде:

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

Это бинарное отношение определяет наличие или отсутствие в речевом слове локального выброса на частоте f в момент времени t. Поскольку для разных произнесений одного и того же слова характерны изменения в структуре локальных выбросов, связанные с изменением интонации, темпа произнесения и т.п. , то для описания эталонного речевого образа необходимо использовать нечёткое бинарное отношение R, которое ставит в соответствие 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 $R = \{r_1, r_2, ... r_n\}$. Каждое нечёткое отношение r_j формируется как среднее арифметическое эталонных ДСВО слова i_j. Входной неизвестный образ y рассматривается как обычное (чёткое) отношение между множеством номеров 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:

$$S_j = \frac{\int \int r_j(f,t) \wedge y(f,t) df dt}{\int \int \neg r_j(f,t) \wedge y(f,t) df dt}.$$

3. Оптимальное выравнивание речевых образов по времени

Различные реализации речевых образов, относящихся к одному и тому же классу, могут значительно отличаться друг от друга по длительности. Это связано с нестабильностью темпа речи диктора, вызванной влиянием интонации, акцента и т.п. Поэтому при вычислении

степени сходства возникает проблема временного выравнивания несовпадающих по длительности входного ДСВО и эталонного нечёткого образа, т.е. приведения их к одинаковой длине по оси времени. Для выполнения этой процедуры в работе [5] был предложен алгоритм линейного выравнивания, заключающийся в приведении образов к заданной длине путём равномерного прореживания или вставок. Недостатком этого алгоритма является то, что он не учитывает неравномерность протекания речевого сигнала во времени. В работе [6] для выравнивания длин сопоставляемых образов применяется нелинейная временная нормализация, реализованная на основе динамического программирования. Однако такой подход требует выполнения длительных вычислений, а также ведёт к снижению различимости речевых образов, относящихся к различным классам. Поэтому в данной работе для приведения сопоставляемых образов к одинаковой длине предложен новый алгоритм оптимального временного выравнивания.

Рассмотрим задачу оптимального временного выравнивания сопоставляемых образов r_j и y разной длительности.

Пусть T_r , T_y – размеры по времени соответственно эталонного образа (нечёткого отношения) $r_j(f, t)$ и входного образа (чёткого отношения) $y(f, t)$, причём $T_r \neq T_y$; F – размер каждого из образов по частоте; $\tilde{r}_j(f,t,h)$ и $\tilde{y}(f,t,h)$ – это образы $r_j(f, t)$ и $y(f, t)$, приведённые к одинаковой длине $T = \max(T_r, T_y)$ по формулам:

$$\tilde{r}_j(f,t,h) = \begin{cases} r(f,t), & T_r > T_y; \\ \hat{r}(f,t,h), & T_r < T_y, \end{cases}$$

$$\tilde{y}(f,t,h) = \begin{cases} \hat{y}(f,t,h), & T_r > T_y; \\ y(f,t), & T_r < T_y, \end{cases}$$

где $\hat{r}(f,t,h)$, $\hat{y}(f,t,h)$ – образы, увеличенные по следующему функциональному правилу:

$$\hat{\varphi}(f,t,h) = \begin{cases} 0, & 1 \le t \le h; \\ \varphi(f,t-h), & h < t \le T - (|T_r - T_y| - h); \\ 0, & T - (|T_r - T_y| - h) < t \le T; \end{cases}$$

$$\hat{\varphi} \in \{\hat{r}, \hat{y}\}; \varphi \in \{r, y\};$$

$$h \in [1, |T_r - T_y|].$$

Задача оптимального временного выравнивания относится к классу задач нелинейной оптимизации функции целочисленного аргумента и имеет вид:

$$S_j(h) = \frac{\sum_{t=1}^{T}\sum_{f=1}^{F} \tilde{r}_j(f,t,h) \wedge \tilde{y}(f,t,h)}{\sum_{t=1}^{T}\sum_{f=1}^{F} \neg\tilde{r}_j(f,t,h) \wedge \tilde{y}(f,t,h)} \longrightarrow \max_h ,$$

$$1 \le h \le |T_r - T_y|.$$

Оптимальное временное выравнивание применяется как при распознавании для попарного выравнивания входного и каждого из эталонных образов, так и при обучении (формировании функций принадлежности).

4. Эксперименты по распознаванию

Ниже описаны эксперименты по однодикторному и многодикторному распознаванию набора русских слов методом нечёткого сопоставления образов с оптимальным выравниванием. Набор слов включал в себя 105 команд управления текстовым редактором, цифр от одного до девяти и обычных слов, записанных 8-битными отсчётами в формате PCM с частотой 11025 Гц. В составлении этого набора принимало участие 9 дикторов (5 мужчин и 3 женщины), каждый из которых по три раза произносил все слова.

При однодикторном распознавании для формирования функций принадлежности использовалось по два из трёх произнесений каждого слова, а для тестирования – по одному произнесению. Эксперименты проводились для каждого из девяти дикторов отдельно, а затем результаты распознавания усреднялись по всем дикторам.

При многодикторном распознавании для формирования функций принадлежности использовалось по два из трёх произнесений каждого слова каждым диктором (итого – 18 произнесений), а для тестирования – по одному произнесению (итого – 9 произнесений). Для исследования зависимости качества распознавания от объёма словаря распознающей системы, кроме полного набора 105 слов, использовались также и сокращённые наборы из 52 и 20 слов.

В экспериментах использовалось 7 различных значений α (порога крутизны локального выброса СВО, используемого для вычисления ДСВО

– см. раздел 1) для того, чтобы определить оптимальное значение α по критерию минимума ошибки распознавания.

Результаты однодикторного распознавания при различных значениях α приведены на рис.2. Установлено, что наибольшая доля правильно распознанных слов – 99,05% – достигается при $\alpha = 179$.

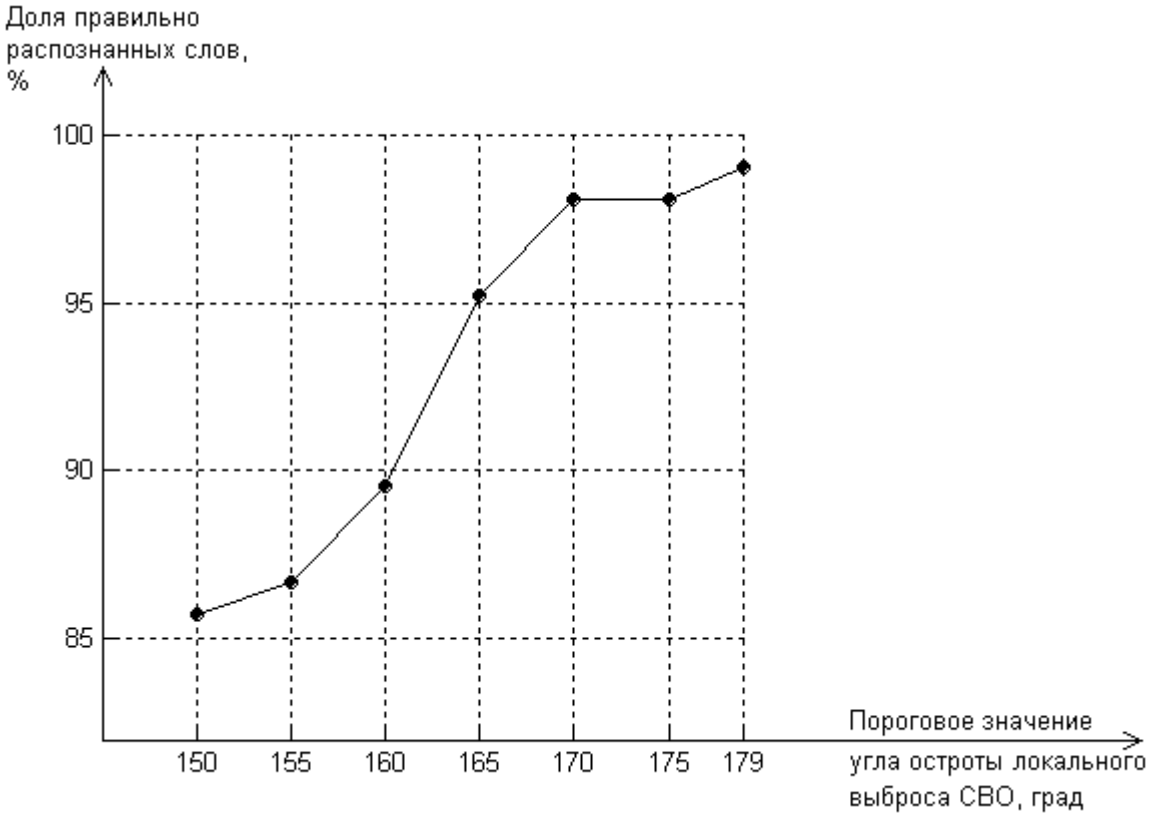

Рисунок 2 – Результаты однодикторного распознавания при различных пороговых значениях угла остроты локального выброса (объём словаря – 105 слов)

Результаты многодикторного распознавания при различных значениях α и различных объёмах словаря распознающей системы приведены на рис.3. Эти результаты свидетельствуют, что наилучшее качество распознавания (81,67% для словаря из 20 слов, 76,28% для словаря из 52 слов и 70,37% для словаря из 105 слов) достигается при значении $\alpha = 179$. Также установлено, что увеличение объёма словаря сопровождается незначительным ухудшением качества многодикторного распознавания для каждого значения α.

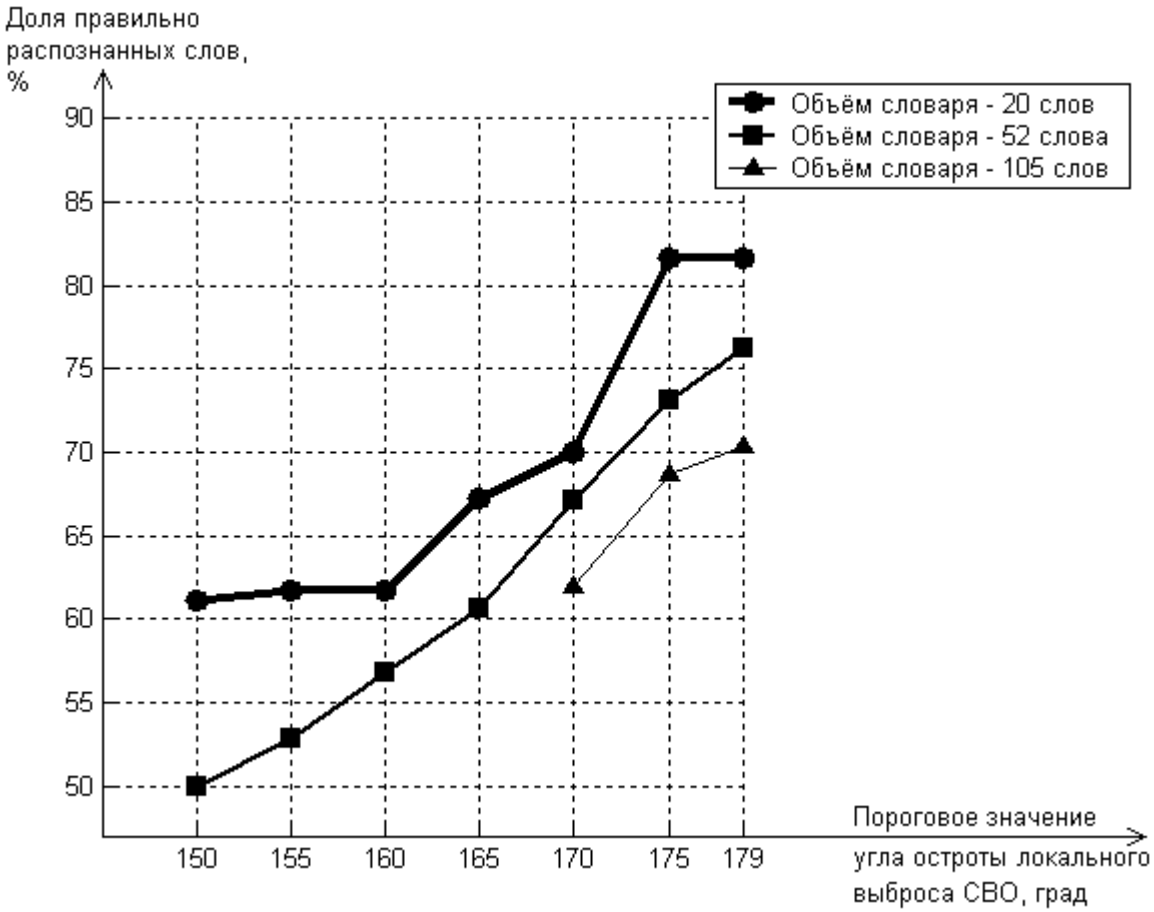

Рисунок 3 – Результаты многодикторного распознавания при различных пороговых значениях угла остроты локального выброса

Заключение

Для распознавания изолированных слов предложен алгоритм нечёткого сопоставления образов с оптимальным временным выравниванием по критерию максимизации степени сходства сопоставляемых образов. Применение этого алгоритма для однодикторного и многодикторного распознавания набора 105 русских слов, образующих команды управления текстовым редактором, показало хороший результат при однодикторном распознавании (99,05% правильно распознанных слов) и менее высокий при многодикторном распознавании (70,37% правильно распознанных слов). С уменьшением объёма словаря распознающей системы до 20 слов качество многодикторного распознавания достигло 81,67%.

Полученные результаты экспериментов свидетельствуют о практической применимости предложенного алгоритма распознавания изолированных слов в системах речевого командного управления с предварительной подстройкой под диктора (пользователя). На основе

предложенного алгоритма возможно также создание систем речевого командного управления и без подстройки под конкретного диктора, но в таких случаях для устойчивого распознавания необходимо использовать словари небольшого размера.

Дальнейшие исследования будут направлены на формирование такого набора признаков речевого сигнала, который повысит инвариантность алгоритма нечёткого сопоставления образов с оптимальным временным выравниванием к изменению голосов дикторов, что позволит создавать системы многодикторного распознавания с большим объёмом словаря.

Перечень ссылок

1. Плотников В.Н. и др. Речевой диалог в системах управления. – М.: Машиностроение, 1988. – 224 с.
2. Кофман А. Введение в теорию нечетких множеств. – М.: Радио и связь. – 1982. – 432 с.
3. Francesco Beritelli, Guiseppe Cilia, Antonino Cucė. Small Vocabulary Word Recognition Based on Fuzzy Pattern Matching // Proc. European Symposium on Intelligent Techniques. – Crete (Greece). – 1999. – http://www.erudit.de/erudit/events/esit99/12651_p.pdf
4. Reza HoseinNezhad, Behzad Moshiri, Parisa Eslambolchi. Fusion of Spectrograph and LPC Analysis for Word Recognition: A New Fuzzy Approach. // Proc. 7th International Conference on Information Fusion. – Stockholm (Sweden). – 2004. – P. 449-454 – http://www.fusion2004.foi.se/papers/IF04-0449.pdf
5. Киедзи Асаи, Дзюндзо Ватада, Сокуке Иваи и др. Распознавание речи // Прикладные нечёткие системы: Пер. с яп. Под ред. Т.Тэрано, К. Асаи, М. Сугено. – М.: Мир, 1993. – с. 157-170.
6. Бондаренко И.Ю., Федяев О.И. Анализ эффективности метода нечёткого сопоставления образов для распознавания изолированных слов // Сб. трудов VI междунар. науч.конференции “Интеллектуальный анализ информации ИАИ-2006”. Под ред. Таран Т.А. – К.: Просвіта, 2006. – с.20–27.