Способы повышения вероятности правильного распознавания для алгоритма распознавания речевых команд по их кросскорреляционным портретам

Е.Ю. Галицкая[1], В.Р. Крашенинников[1]

[1]Ульяновский государственный технический университет, Северный Венец 32, Ульяновск, Россия, 432027

Аннотация. В настоящее время интенсивно расширяется область применения речевых информационно-управляющих систем (РИУС), для чего необходимо распознавание речевых команд (РК). Это распознавание очень затруднено при наличии сильных акустических помех. В данной работе рассматривается метод распознавания сильно зашумленных РК по кросскорреляционным портретам (ККП), который используется при дикторозависимом распознавании из ограниченного словаря команд. В этом методе РК преобразуются в ККП, представляющие собой особые изображения, отражающие особенности акустического звучания команды. От выбора эталонов команд напрямую зависит вероятность правильного распознавания. Эталоны должны отражать достаточно точно весь класс команд, для чего производится оптимизация библиотеки эталонов. В памяти компьютера эталоны хранятся в виде ККП. Распознаваемая РК также преобразуется в портрет и находится наиболее близкий портрет из множества портретов эталонов. При этом требуется достаточно точное совмещение портретов эталона и распознаваемой РК. Для этого предлагаются два способа уточнения совмещения: пофонемное совмещение и варьирование границ РК, учитывая, что её границы могут быть оценены с опережением или запаздыванием. Проведенные эксперименты показали, что предлагаемая модернизация алгоритма существенно повышает вероятность правильного распознавания РК.

1. Введение

Несмотря на значительные успехи роботизации, в настоящее время управление многими техническими системами невозможно без участия человека-оператора. При этом желательно снизить нагрузку на оператора, что может быть достигнуто применением речевых информационно-управляющих систем (РИУС), в которых возможно получение информации о состоянии системы и управление ею по голосовым запросам и командам, для чего необходимо распознавание речевых команд (РК). К настоящему времени разработано множество систем распознавания речи, применяемых для ввода информации в компьютер, управления роботами и т.д. [1-5]. Однако большинство таких систем работоспособно при отсутствии шумов (или слабых). В то же время имеется необходимость в РИУС, работающих в условиях очень сильных акустических помех, например, шумные производства авиация и т.д. Исследования по созданию таких систем ведутся. Например, системы распознавания РК пилота военного самолёта [6-10]. Однако ни одна из них не была успешно внедрена по причине низкой вероятности правильного распознавания РК в условиях особо сильных акустических помех.

Таким образом, остаётся актуальной задача создания методов и алгоритмов распознавания РК в условиях сильных помех.

Обычно системы распознавания РК в условиях сильных помех являются дикторозависимыми с ограниченным словарём. Для РК из этого словаря строятся некоторые эталоны, отражающие существенные признаки команд, а распознаваемая РК относится к наиболее близкому из этих эталонов. В данной работе рассматривается метод распознавания сильно зашумленных РК по кросскорреляционным портретам (ККП). В этом методе РК преобразуются в ККП, представляющие собой особые изображения, отражающие особенности акустического звучания команды. В памяти компьютера эталоны хранятся в виде ККП. Распознаваемая РК также преобразуется в портрет и находится наиболее близкий портрет из множества портретов-эталонов. От выбора эталонов команд существенно зависит вероятность правильного распознавания, поэтому производится оптимизация библиотеки эталонов. При этом требуется достаточно точное совмещение портретов эталона и распознаваемой РК. Для этого предлагаются два способа уточнения совмещения: пофонемное совмещение и варьирование границ РК, учитывая, что её границы могут быть оценены с опережением или запаздыванием. Проведенные эксперименты показали, что предлагаемая модернизация метода существенно повышает вероятность правильного

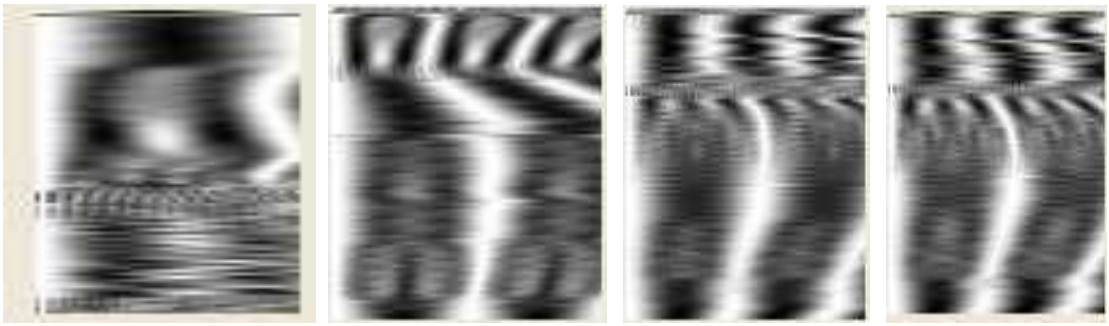
2. Распознавание речевых команд по их автокорреляционным и кросскорреляционным портретам

В [11] для распознавания РК на фоне сильных шумов было предложено использование их автокорреляционных портретов (АКП). Пусть $X=\{x_1,x_2,...,x_N\}-$ РК, состоящая из N отсчетов. Построим её АКП, являющееся двумерным массивом (изображением). Для этого разобьем X на $M+1$ равных отрезков $X_1,...,X_{M+1}$ длины $L=[N/(M+1)]$, где [ ] – целая часть числа. Каждая строка АКП представляет собой последовательность выборочных коэффициентов корреляции $r(t,k)$ t-го отрезка $X_t=\{x_{(t-1)L+1},...,x_{tL}\}$ и сдвинутых на k отсчетов $X_t=\{x_{(t-1)L+1+k},...,x_{tL+k}\}$, где k=0…K:

$$r(t,k)=\frac{1}{L\sigma_t\sigma_{t+k}}\left(\sum_{j=0}^{L-1}x_{(t-1)L+j}x_{(t-1)L+j+k}-\mu_t\mu_{t+k}\right), \qquad (1)$$

где $t=1,..,M$, $k=1,...,K$, μ_t и μ_{t+k} – выборочные средние, σ_t^2 и σ_{t+k}^2 – выборочные дисперсии. Таким образом, АКП является массивом (изображением) размеров $M\times K$ из выборочных коэффициентов автокорреляции одной РК X. На рисунке 1 показаны АКП РК «Кабина», «Двигатель» и двух произнесений «Кондиционер» в разное время. Здесь диапазон значений [-1; 1] коэффициента корреляции преобразован в диапазон яркостей [0; 256]. Строка изображения отражает изменение коэффициента корреляции между значениями речевого сигнала при сдвигах на $k=1,2,...,K$ отсчётов, то есть локальные связи, например, внутрифонемные. Последовательность строк отражает процесс изменения корреляций со временем звучания команды, например, характеризует последовательность фонем.

Оказалось, что АКП достаточно индивидуальны для распознавания РК по их портретам, довольно устойчивы к шумам, слабо чувствительны к громкости произнесения. Главным достоинством является сильная межстрочная коррелированность портретов, что даёт возможность применения методов обработки изображений при решении задач фильтрации, распознавания и т.д. Однако имеется и существенный недостаток – АКП отражает особенности одного произнесения РК, по которому он и построен. Это заметно по двум портретам команды «Кондиционер», построенным из произнесений, полученным в разное время. За это время тембр голоса диктора, темп речи, состояние здоровья и т.д. могли существенно измениться. Эталоны как бы «старели», поэтому портрет эталона и портреты той же распознаваемой команды могли значительно различаться, что снижало качество распознавания. Поэтому эталоны нужно время от времени обновлять.



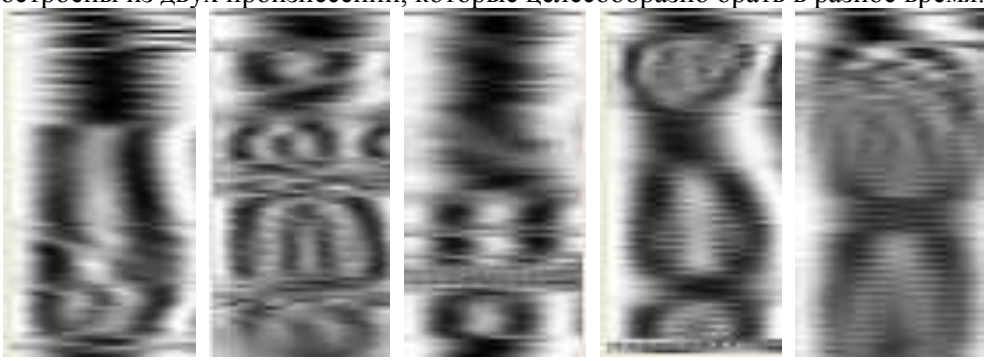
«Борт» «Двигатель» «Кондиционер1» «Кондиционер2»

Рисунок 1. Примеры автокорреляционных портретов речевых команд.

Полнее свойства РК представлены в её ККП, которое строятся по двум произнесениям [12]. Пусть X и Y – два произнесения одной и той же РК одним диктором в разное время. Они разбиваются на одинаковое количество M отрезков с длинами L_X и L_Y соответственно. Каждая строка ККП представляет собой последовательность выборочных коэффициентов корреляции $r(t,k)$ t-го отрезка РК X со сдвинутыми на k отсчётами t-го отрезка РК Y:

$$r(t,k) = \frac{1}{L_X \sigma_{X,t} \sigma_{Y,t+k}} \left(\sum_{j=0}^{L_X - 1} x_{(t-1)L_X + j} y_{(t-1)L_Y + j + k} - \mu_{X,t} \mu_{Y,t+k} \right), \quad (2)$$

где $t=1,..,M$, $k=1,...,K$, $\mu_{X,t}$ и $\mu_{Y,t+k}$ – выборочные средние, $\sigma^2_{X,t}$ и $\sigma^2_{Y,t+k}$ – соответствующие выборочные дисперсии. Таким образом, ККП является массивом (изображением) размеров $M \times K$ из выборочных коэффициентов кросскорреляции двух РК X и Y. Если $X=Y$, то ККП совпадает с АКП. На рисунке 2 показаны ККП шести РК из двух их произнесений при количестве отрезков разбиения (то есть строк) $M=100$ и количестве сдвигов (то есть столбцов) $K=50$. Заметно, что ККП различных команд индивидуальны, что делает их хорошей основой для распознавания. В то же время, они в большей мере отражают вариативность произношения, так как построены из двух произнесений, которые целесообразно брать в разное время.



«Борт» «Дозаправка» «Двигатель» «Маневр» «Питание»

Рисунок 2. Примеры кросскорреляционных портретов речевых команд.

В памяти компьютера эталоны хранятся в виде ККП. Распознаваемая РК также преобразуется в ККП в паре с каким-нибудь заранее начитанным произнесением, например, из эталонных. По некоторой метрике находится наиболее близкий к ней портрет из множества портретов-эталонов. При этом расстояние между двумя ККП определяется как сумма расстояний между соответствующими строками в некоторой метрике, например, евклидовой. Сравниваемые команды могут быть произнесены в разном темпе, поэтому нужно определить соответствие между строками их портретов, для чего применялось динамическое программирование.

3. Методы повышения вероятности правильного распознавания команд

При отсутствии шумов описанный метод распознавания даёт практически стопроцентную вероятность правильного распознавания. Наличие сильных шумов значительно её снижает по ряду причин. Рассмотрим некоторые из мешающих факторов и снижения их влияния. Некоторые из этих методов были применены для улучшения распознавания команд по их АКП [13, 14].

Варьирование границ распознаваемой команды. Для сравнения эталонов с распознаваемой командой прежде всего требуется определить её начало и конец. При этом из-за сильных шумов неизбежны ошибки – опережение или запаздывание. Особенно трудно найти конец команды, так как он обычно произносится тише, чем начало. Для ослабления влияния этих ошибок были применены пробные добавления и удаления нескольких отсчётов сигнала на оцененных границах. Из перебранных вариантов выбирался тот, который давал наименьшее расстояние.

Оптимизация параметра *K*. Параметр K равен ширине ККП и выбирается эмпирически. Однако, как показала практика, оптимальное значение параметра K зависит от длины РК. Поэтому все команды словаря были разбиты на группы примерно одинаковой длины, и для каждой группы использовалось своё значение этого параметра.

Совмещение фонем при построении ККП. При построение ККП команды разбиваются на N отрезков. В каждый отрезок попадает какая-то часть фонемы. Из-за изменчивости темпа произнесения, начала отрезков команд могут начинаться с различных фонем, поэтому коэффициент корреляции может иметь «ложное» значение и ККП будет искажаться. Для борьбы с данным искажением использовался алгоритм динамического совмещения фонем. В результате начало отрезка одной команды сдвигается так, чтобы этот отрезок максимально коррелировался с отрезком второй команды.

Оптимизация библиотеки эталонов. Качество распознавания напрямую зависит от того, насколько хорошо эталонные ККП отписывают варианты произнесения команд. В связи с этим возникает дополнительная задача выбора «наилучших» эталонов. Для этого сначала строится по нескольку эталонов каждой команды и направленным перебором выбирается по одному эталону каждой команды для достижения наилучшего качества распознавания большого набора произнесений команд. Для выполнения этой операции желательно иметь большое количество произнесений РК, что требует больших временных затрат дикторов. В [15, 16] описаны способы получения реализаций квазипериодических процессов в виде авторегрессионных моделей цилиндрических изображений. Фонемы речевых сигналов тоже являются квазипериодическими процессами, что позволило моделировать множество вариантов произнесения РК даже из одного её реального произнесения диктором.

Зашумление эталонов. Эталоны строятся обычно заранее по незашумлённым произнесениям. Распознаваемая же РК содержит значительный шум, поэтому её ККП неизбежно отличается от эталонного ККП, поэтому расстояния между ККП искажаются и снижается качество распознавания. Для коррекции расстояний было применено зашумление эталонных РК перед их преобразованием в ККП. При этом шум для эталонов поступал с дополнительного микрофона вдали от рта оператора во время произнесения распознаваемой команды, что обеспечивало близость характеристик шума в сравниваемых ККП. Недостатком этого способа является вычисление всех зашумлённых эталонов при распознавании каждой поступающей команды.

4. Результаты экспериментов

Для оценки значимости рассмотренных способов повышения вероятность правильного распознавания был проведен следующий эксперимент. Имелся словарь, состоящий из 10 РК авиационной тематики. Каждая РК была произнесена 20 раз (всего в распознавании участвовало 200 РК). Команды были аддитивно зашумлены шумом авиационного двигателя с отношением сигнал/шум 4. При обычном распознавание в качестве эталонных произнесений были выбраны первые два произнесения. В результате распознавания была выделена группа 34 команд, которые не были распознаны. Применение рассмотренных выше методов позволило

распознать 16 из нераспознанных команд. При этом команды, распознанные верно в первом случае, были также распознаны верно улучшенным методом. В результате вероятность правильного распознавания повысилась с 83% до 91% (значимость проверена по критерию Стьюдента с уровнем значимости 0.05).

5. Заключение

Для распознавания РК на фоне сильных шумов предложено использование преобразование РК в ККП, то есть в двумерные изображения, строки которых состоят из коэффициентов кросскорреляциии между двумя произнесениями команды. Использование двух произнесений в портрете позволяет в некоторой мере учесть вариативность произнесения. Распознавание осуществляется сравнением портреты распознаваемой команды с эталонными портретами. Эксперименты показали, что применение нескольких модификаций этого метода существенно увеличивает вероятность правильного распознавания.

6. Благодарности

Исследование выполнено при финансовой поддержке РФФИ, проект 20-01-00613.

7. Литература

[1] Жданов, А. Речевой ввод как альтернатива клавиатурному – [Электронный ресурс]. – Режим доступа: https://compress.ru/article.aspx?id=11907 (05.11.2019).

[2] Михайлюк, М.В. Эргономичный голосовой интерфейс управления антропоморфным роботом – [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/ergonomichnyy-golosovoy-interfeys-upravleniya-antropomorfnym-robotom/viewer (05.11.2019).

[3] Голосовое управление, роботы и энергия из воздуха. Какие технологии изменят мир в ближайшем будущем – [Электронный ресурс]. – Режим доступа: https://rus.delfi.lv/techlife/obzory/golosovoe-upravlenie-roboty-i-energiya-iz-vozduha-kakie-tehnologii-izmenyat-mir-v-blizhajshem-buduschem.d?id=47144161&all=true (05.11.2019).

[4] Умный дом от Apple, Google и Яндекс – голосовое управление – [Электронный ресурс]. Режим доступа: https://voiceapp.ru/articles/smarthome (05.11.2019).

[5] SpeechKit – речевые технологии Яндекса – [Электронный ресурс]. Режим доступа: https://yandex.ru/company/technologies/speech_technologies/ (05.11.2019).

[6] Суворов, Ф.А. Внедрение Voice Flight VFS101 в управление навигационной системой на борту самолета гражданской авиации – [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/vnedrenie-voice-flight-vfs101-v-upravlenie-navigatsionnoy-sistemoy-na-bortu-samoletov-grazhdanskoy-aviatsii/viewer (05.11.2019).

[7] Создание Euro-canards – [Электронный ресурс]. Режим доступа: http://top-airplane.ru/sozdanie_euro_canards.html (05.11.2019).

[8] «Ратник» получит голосовую систему управления – [Электронный ресурс]. Режим доступа: https://topwar.ru/98285-ratnik-poluchit-golosovuyu-sistemu-upravleniya.html (05.11.2019).

[9] «Матрица» в F-35: шлем с дополненной реальностью для пилотов – [Электронный ресурс]. Режим доступа: https://www.popmech.ru/weapon/295312-matritsa-v-f-35-shlem-s-dopolnennoy-realnostyu-dlya-pilotov/ (05.11.2019).

[10] Кучерявый, А.А. Бортовые информационные системы. – [Электронный ресурс]. Режим доступа: http://window.edu.ru/catalog/pdf2txt/082/59082/29039?p_page=4 (05.11.2019).

[11] Крашенинников, В.Р. Распознавание речевых команд на фоне интенсивных помех с помощью авторегрессионных портретов / В.Р. Крашенинников, А.И. Армер, Н.А. Крашенинникова, А.В. Хвостов // Наукоемкие технологии. – 2007. – № 9. – С. 65-74.

[12] Krasheninnikov, V.R. Cross-Correlation Portraits of Voice Signals in the Problem of Recognizing Voice Commands According to Patterns / V.R. Krasheninnikov, A.I. Armer, V.V. Kuznetsov, E.Yu. Lebedeva // Pattern Recognition and Image Analysis. – 2011. – Vol. 21(2). – P. 185-187.

[13] Krasheninnikov, V.R. Optimization of dictionary and model library for recognition of speech commands / V.R. Krasheninnikov, V.V. Kuznetsov, E.Yu. Lebedeva // Pattern Recognition and Image Analysis. – 2011. – Vol. 21(3). – P. 505-507.
[14] Krasheninnikov, V.R. Preparation of Templates in Speech Command Recognition by Single- and Double-Channel Scheme in Background Noise / V.R. Krasheninnikov, A.V. Khvostov, A.I. Armer // Pattern Recognition and Image Analysis. – 2008. – Vol. 18(4). – P. 580-583.
[15] Krasheninnikov, V.R. Autoregressive Models of Speech Signal Variability in the Speech Commands Statistical Distinction / V.R. Krasheninnikov, A.I. Armer, N.A. Krasheninnikova, V.P. Derevyankin, V.I. Kozhevnikov, N.N. Makarov // Internetional Conference on Computational Science and it's Applications – Springer-Verlag: Berlin Heidelberg, 2006. – P. 974-982.
[16] Krasheninnikov, V.R. Multidimensional Image Models and Processing / V.R. Krasheninnikov, K.K. Vasil'ev // Computer Vision in Control Systems-3. Intelligent Systems Reference Library – Springer International Publishing, 2018. – P. 11-64.

Ways to increase the probability of correct recognition of noisy speech commands by their cross-correlation portraits

E.Yu Galitskaya[1], V.R. Krasheninnikov[1]

[1]Ulyanovsk State Technical University, Severny Venets street 32, Ulyanovsk, Russia, 432027

Abstract. Currently, the field of application of voice information-control systems is being intensively expanded, for which recognition of speech commands (SC) is necessary. This recognition is very difficult in the presence of intense acoustic noise. We consider a method for recognizing noisy SCs by cross-correlation portraits (CCP), which is used for speaker-dependent recognition from a limited vocabulary of commands. In this method, SCs are converted to CCPs, which are special images. The probability of correct recognition directly depends on the choice of command standards. The standards should accurately reflect the entire class of commands, for which the library of standards is optimized. The standards are stored as CCPs. Recognized SC is converted into CCP and the closest portrait is found from the set of portraits of standards. In this case, a sufficiently accurate coincidence of the portraits of the standard and the recognizable SC is required. For this, two methods are proposed: phonemic alignment and variation of the boundaries of SC, given that its boundaries can be estimated ahead or delayed. The experiments showed that the proposed modernization of the algorithm significantly increases the probability of correct recognition.