УДК 004.522:004.934

ОРГАНИЗАЦИЯ СИСТЕМЫ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ НА ОСНОВЕ КОЛЛЕКТИВА РАСПОЗНАЮЩИХ АВТОМАТОВ

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

У даній статті за допомогою апарату теорії інформації показана принципова доцільність створення системи автоматичного розпізнавання усного мовлення на основі методів колективного розпізнавання. Проаналізовані різні засоби організації окремих автоматів, що розпізнають, у колектив. Розроблена обчислювальна структура системи розпізнавання фонем злитого мовлення, у якій для формування колективу застосовується метод bagging, а для реалізацій автоматов, що розпізнають, використовуються штучні нейронні мережі. Виконані експериментальні дослідження, що спрямовані на оцінювання точності роботи такої системи.

Введение

Устная речь является наиболее естественным для человека способом общения. Использование устной речи в диалоге с компьютерами, роботами, автоматизированными системами управления с помощью речевых сообщений открывает большие перспективы:

1) простота общения с системой (использование речевого интерфейса не требует специальной подготовки оператора, т.к. общение происходит на естественном языке);

2) доступность речевого интерфейса людям с нарушениями опорно-двигательного и зрительного аппарата;

3) возможность работы пользователей в условиях перегруженности тактильно-зрительных каналов.

Для построения систем речевого диалога и управления необходимо решить задачи автоматического распознавания и синтеза устной речи. Если проблема речевого синтеза на научном уровне уже решена, и продолжаются лишь разработки, направленные на точное воспроизведение индивидуальных особенностей человеческих голосов [2, 3], то проблема распознавания речи, несмотря на

множество предложенных подходов к её решению, по-прежнему остаётся открытой. Так, существующие в 2004 году разработки систем распознавания речи обеспечивали приемлемую точность 90-95% распознанных слов лишь при длительной подстройке под конкретного диктора и чётком произнесении отдельных слов и словосочетаний [4], и сейчас ситуация к лучшему практически не изменилась.

Таким образом, в настоящее время создание достаточно точной системы распознавания устной речи, инвариантной к изменению дикторов, по-прежнему остаётся актуальной научной задачей.

Целью данной работы является повышение точности функционирования системы автоматического распознавания речи за счёт применения методов коллективного распознавания образов. *Объектом исследования* в данной работе является система автоматического распознавания речи, а *предметом* — методы и алгоритмы формирования коллектива речевых распознавателей.

1 Информационное обоснование повышения достоверности коллективного распознавания

Ввиду того, что речь представляет собой нелинейный нестационарный процесс, не удаётся устойчиво выделить признаки речевых образов, которые позволяют проводить абсолютно безошибочную классификацию этих образов в системе автоматического распознавания. Информативность в отдельности используемых в настоящее время признаков речи (спектр, кепстр, коэффициенты линейного предсказания и т.п.) мала, поскольку эти признаки не являются достаточно инвариантными к искажениям, связанным с индивидуальными особенностями и эмоциональным состоянием говорящего, а также к нестационарным фоновым шумам. В данной ситуации повышение достоверности распознавания речи только за счёт увеличения числа неустойчиво выделяемых признаков и связанного с этим увеличения априорной информации не сможет гарантировать хорошее распознавание.

В этих условиях для повышения достоверности распознавания речевых образов логично увеличить количество текущей информации о распознаваемом образе за счёт объединения отдельных распознавателей в единую систему на принципах коллективного распознавания речи.

Все N распознавателей в такой системе одновременно проводят классификацию поступившего на вход речевого образа. При этом, если разные распознаватели используют один и тот же набор признаков X (например, только спектрограмму речевого сигнала), как

это показано на рис.1а, то объединение их в систему приводит к увеличению количества текущей информации. Если же наборы признаков X_i , i = 1..N, индивидуальны для каждого из распознавателей (например, первый распознаватель использует спектрограмму речевого сигнала, второй — мел-частотную кепстрограмму, а третий — набор вейвлет-коэффициентов), как это показано на рис.1б, то в результате объединения таких распознавателей в систему увеличивается количество как текущей, так и априорной информации.

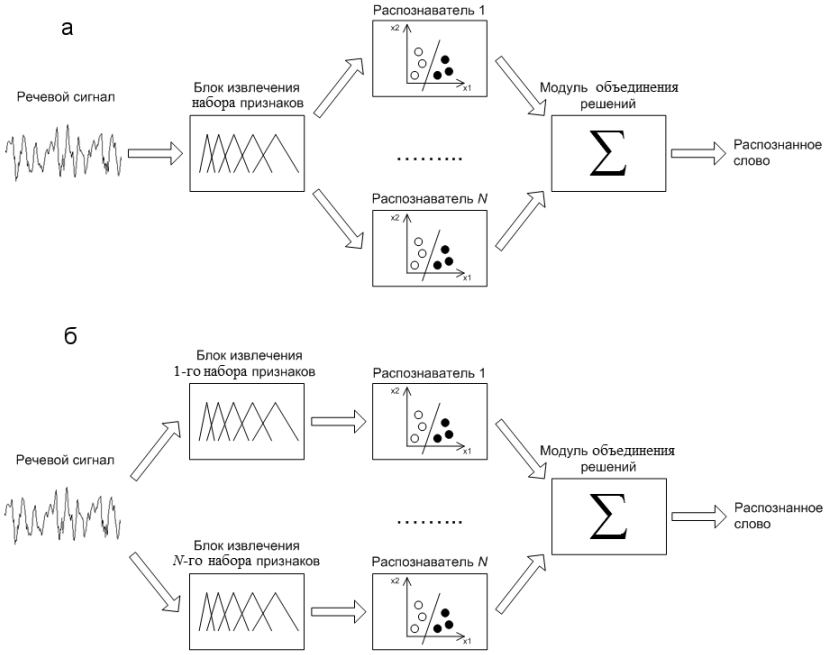

Рисунок 1 – Объединение отдельных распознавателей в систему: а — все распознаватели используют один и тот же набор признаков речевого сигнала; б — распознаватели используют разные наборы признаков речевого сигнала

Согласно теории информации [5], как в том, так и в другом случае количество информации, перерабатываемой системой при коллективном распознавании, составляет:

$$I(A, X) = H(A) - H(A|X) \qquad (1)$$

где $A = \{A_i\}$, i = 1..M – словарь распознавания (множество распознаваемых слов или фонем); $X = \{X_j\}$, j = 1..N – множество наборов признаков, используемых распознавателями; $H(A)$ – энтропия на входе системы распознавания (исходная энтропия), вычисляемая по формуле:

$$H(A) = -\sum_{i=1}^{M} P(A_i) \cdot \log(P(A_i)); \quad (2)$$

$H(A|X)$ – энтропия на выходе системы распознавания (энтропия решения), которая для коллектива из N распознающих автоматов вычисляется по формуле:

$$H(A \mid X) = H(A \mid X_{1,}...,X_N). \quad (3)$$

Энтропия решения системы коллективного распознавания $H(A|X)$ может уменьшаться при увеличении числа распознавателей в системе, поскольку, согласно [6], условная энтропия с ростом числа фиксируемых условий не возрастает, т. е.

$$H(A \mid X_{1,}...,X_N) \le H(A \mid X_{1,}...,X_{N-1}). \quad (4)$$

При этом строгое равенство имеет место тогда и только тогда, когда выполняется условие:

$$p(A, X_N \mid X_{1,}...,X_{N-1}) = p(A \mid X_{1,}...,X_{N-1}) \times p(X_N \mid X_{1,}...,X_{N-1}). \quad (5)$$

Это означает, что при распознавании речи данные, предоставляемые соседними распознавателями коллектива, не дают дополнительной информации к той, которой располагает каждый конкретный распознаватель. Это возможно в двух случаях: 1) один из распознавателей всегда принимает безошибочные решения; 2) все распознаватели всегда принимают одинаковые решения. Первый случай невозможен на практике, а второй случай исключается путём формирования коллектива из разнообразных, а не одинаковых, распознающих автоматов.

2 Способы объединения распознающих автоматов в коллектив

Таким образом, возникает задача определения структуры коллектива распознавателей. Существует ряд методов формирования коллектива разнообразных распознающих автоматов [7], среди которых можно выделить три основных:

1) bagging, или bootstrap aggregation – обучение распознающих автоматов на бутстрап-подмножествах исходного обучающего множества [8];

2) boosting – последовательное обучение распознающих автоматов коллектива, при котором каждый следующий распознающий автомат, включаемый в коллектив, обучается так, чтобы компенсировать недостатки всех предыдущих автоматов [9];

3) mixture of experts – смесь экспертов, когда в коллектив вводится дополнительный распознающий автомат, оценивающий компетентность остальных членов коллектива для каждого входного образа и объединяющий индивидуальные решения с учётом этих оценок [10].

Задача распознавания речи характеризуется высокой вычислительной сложностью и большими объёмами обучающих данных (например, классическая речевая база данных для обучения распознаванию английской речи TIMIT [11] содержит свыше 500 Мб речевого материала). Для решения такой задачи наиболее целесообразным представляется использование первого подхода – формирования коллектива на основе метода bagging, потому что:

1) обучение отдельных распознающих автоматов на своих обучающих бутстрап-подмножествах происходит независимо, что позволяет ускорить формирование коллектива за счёт распараллеливания процессов обучения отдельных распознающих автоматов;

2) обучающее бутстрап-подмножество может быть меньше, чем исходное обучающее множество, что позволяет ускорить процесс обучения каждого распознающего автомата.

Для контекстно-независимого распознавания фонем в слитной речи авторами предложена следующая вычислительная структура (см. рис.2), основанная на применении bagging-коллектива распознающих автоматов [12].

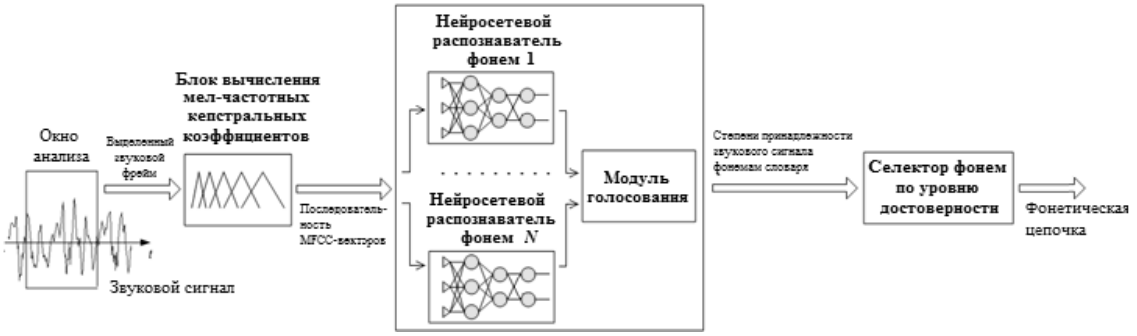

Рисунок 2 – Структура системы распознавания фонем устной речи на основе bagging-коллектива нейросетевых распознавателей

Здесь в качестве автоматов применены многослойные нейронные сети с сигмоидальными функциями активации, обучаемые по методу обратного распространения ошибки [13]. Решение членов такого коллектива объединяются путём равноправного голосования, а формируется он с помощью вышеупомянутого метода bagging.

3 Результаты экспериментов и их обсуждение

Были проведены эксперименты для сравнения точности распознавания фонем в речевом сигнале системой на основе bagging-коллектива нейросетевых распознавателей (рис. 2) и системой на основе одиночного нейросетевого распознавателя (рис. 3).

Рисунок 3 – Структура системы распознавания фонем устной речи на основе одиночного нейросетевого распознавателя

В качестве материала для экспериментов использовалась классическая речевая база TIMIT, содержащая более 5 часов звукозаписей различных английских фраз, которые были произнесены 630 дикторами на 8 диалектах американского английского языка. Все звукозаписи имеют временную пофонемную маркировку, выполненную профессиональными фонетистами. Речевая база разбита на два непересекающихся множества: обучающее и тестовое [11].

В ходе экспериментов одиночный нейросетевой распознаватель, который был обучен на всём обучающем множестве, выполнил распознавание с точностью 61,13%. Bagging-коллектив из 50 нейронных сетей, каждая из которых обучалась на bootstrap-подмножестве объёмом 40% от объёма исходного обучающего множества, показал более высокую точность распознавания – 63,72%. Зависимость точности распознавания от размера bagging-коллектива показана на рис. 4.

Вышеприведённые результаты экспериментов соответствуют уровню известных зарубежных исследовательских центров речевых технологий (например, от 52% до 64% правильно распознанных фонем на материале той же речевой базы TIMIT в работах [14-16]).

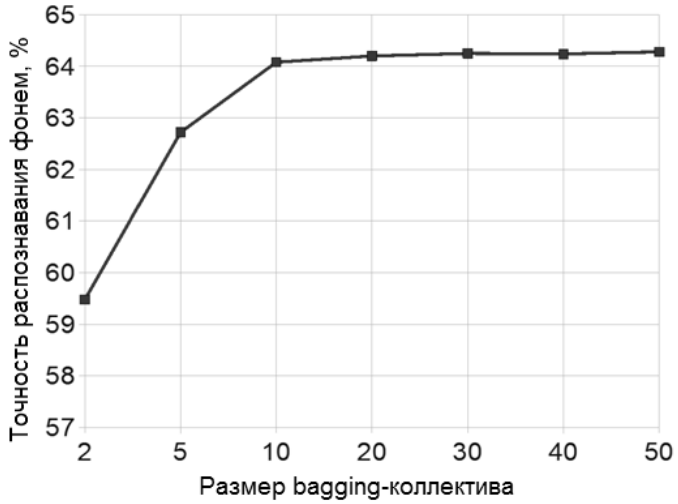

Рисунок 4 – Зависимость точности распознавания фонем от размера bagging-коллектива нейросетевых распознавателей

В дальнейшем планируется разработка лингвистического блока системы распознавания, который позволит на основе распознаваемых фонемных цепочек формировать целые слова и фразы.

Выводы

В работе предложено решение проблемы автоматического распознавания речи с помощью методов коллективного распознавания. Показана теоретическая целесообразность организации системы автоматического распознавания речи на основе коллектива распознающих автоматов. Проведён сравнительный анализ различных методов формирования такого коллектива. В результате анализа сделан вывод о том, что лучшим вариантом для системы распознавания речи по критерию качество/ресурсоёмкость является формирование коллектива распознавателей на основе метода Bootrstap Aggregation, или bagging.

Разработана вычислительная структура системы контекстно-независимого распознавания фонем устной речи, основанная на объединении отдельных нейронных сетей, распознающих фонемы, в bagging-коллектив. На материале классической речевой базы TIMIT проведены экспериментальные исследования, продемонстрировавшие высокую эффективность bagging-коллектива нейросетевых распознавателей фонем по сравнению с другими подходами к распознаванию.

Список литературы

1. Гладун В.П. Партнёрство с компьютером. – К.: «Port-Royal», 2000. – 128 с.

2. Вінцюк Т.К., Сажок М.М., Людовик Т.В., Селюх Р.А. Автоматичний озвучувач українських текстів на основі фонемно-трифонної моделі з використанням природного мовного сигналу // Праці 6-ї міжнародної конференції «УкрОбраз-2002». – К.: Видання Міжнар. науково-навчального центру інф. технологій та с-м, 2002. – С.79-84.

3. Лобанов Б.М., Цирульник Л.И. Компьютерный синтез и клонирование речи. – Минск: Белорусская Наука, 2008. – 316 с.

4. V.I.Galunov, A.N.Soloviev, V.K.Uvarov. Models of Speech Perception, Speech Production and Problem Automatic Speech Recognitions // Proceedings of International Conference on Speech and Computer SPECOM-2004. – Saint-Petersburg, 2004.

5. Барабаш Ю.Л. Коллективные статистические решения при распознавании. – М.: Радио и связь, 1983. – 224 с.

6. Файнстейн А. Основы теории информации. Пер. с англ. – М.: Изд-во иностранной лит-ры, 1960. – 143 с.

7. Городецкий В.И., Серебряков С.В. Методы и алгоритмы коллективного распознавания // Автоматика и телемеханика, №11. – 2008. – С. 3-40.

8. Breiman L. Bagging predictors // Machine Learning. – 1996. – Vol.24, №2. – P. 123-140.

9. D. L. Shrestha, D. P. Solomatine. Experiments with AdaBoost.RT, an Improved Boosting Scheme for Regression // Neural Computation, Vol. 18, No. 7. – 2006. – P.1678-1710.

10. Ran Avnimelech, Nathan Intrator. Boosted Mixture of Experts: An Ensemble Learning Scheme // Neural Computation, Vol. 11, No. 2. – 1999. – P. 483-497.

11. Zue V., Seneff S., Glass J. Speech database development at MIT: TIMIT and beyond // Speech Communication. – 1990. – Vol. 9, № 4. – P.351-356.

12. Федяев О.И., Бондаренко И.Ю. Сегментация речевого сигнала на основе bagging-коллектива нейросетевых детекторов фонем // Материалы 8-й Международной научно-практической конференции «Математическое и программное обеспечение интеллектуальных систем» MPZIS-2010. – Днепропетровск: ДНУ. – 2010. – С.238-239

13. LeCun Y., Bottou L., Orr G., Muller K. Efficient BackProp // Neural Networks: Tricks of the trade. – Springer Verlag, 1998. – P. 5-50.

14. Lee K.-F., Hon H.-W. Speaker Independent Phone Recognition Using Hidden Markov Models // IEEE Transactions on Acoustics, Speech and Signal Processing. – 1989. – Vol.37, №11 – P. 1641-1648.

15. Glass J., Chang J., McCandless M. A probabilistic framework for feature-based speech recognition // Fourth International Conference on Spocen Language «ICSLP 96» Proceedings. – 1996. – Vol.4 – P. 2277-2280.

16. Becchetti C., Ricotti L.P. Speech Recognition: Theory and C++ Implementation. – John Wiley & Sons, 1999. - 428 p.

Получено 12.09.2011