Компьютерные технологии текстонезависимой верификации дикторов на основе алгоритма самоорганизации

Eugene Sidorenko

1. ВВЕДЕНИЕ

Большие объемы данных, накопленные в телекоммуникационных системах, могут содержать информацию о тексте высказываний, о личности говорящего, его эмоциональном состоянии, состоянии здоровья и прочих дополнительных факторах. Обработка этой информации в биометрических системах позволяет получать новые знания о личности говорящего, проводить семантический анализ высказываний, автоматически анализировать диалоги. Биометрические системы находят применение при автоматической идентификации в телефонии, при обработке речевых баз данных, в криминалистических исследованиях, в системах контроля и управления доступом к физическим объектам и к удаленным ресурсам в компьютерных сетях. Биометрические системы имеют перспективу применения в производстве при контроле доступа к оборудованию.

Существующие биометрические системы, выполняющие идентификацию, верификацию диктора,основаны на формальных методах обработки сигналов, и, к сожалению, имеют значительные ошибки принимаемых решений. Ошибки вызваны существенной изменчивостью речи, зависят от типа телекоммуникационных каналов, метода кодирования, алгоритма сжатия, возрастают в условиях внешних акустических шумов. В то же время человек практически без ошибок распознает голос знакомого диктора при различных внешних условиях и разном качестве речи. Поэтому повышение качества работы биометрических систем целесообразно связывать с разработкой биологически обоснованных методов и алгоритмов обработки речи, моделирующих органы слуха человека.

По методу решения идентификации и верификации диктора делятся на текстозависимые и текстонезависимые, а так же работающие на открытом или закрытом множестве дикторов. Если входной сигнал гарантированно принадлежит одному из множества кандидатов, то распознавание происходит на закрытом множестве дикторов. Если сигнал может не принадлежать ни одному из кандидатов, то задача распознавания решается на открытом множестве дикторов.

Распознавание является текстозависимым, если алгоритмы обучены на определенной фразе, которую должен произносить диктор. Распознавание дикторов является текстонезависимым, если предполагается, что диктор будет произносить произвольную фразу. Эффективность текстонезависимого распознавания обычно ниже, чем текстозависимого, однако, сфера применения подобных алгоритмов более широкая, поскольку диктор может не взаимодействовать с системой распознавания и не обязан произносить определенную фразу.

Существующие биометрические методы текстонезависимой идентификации и верификации дикторов основанына цифровой обработке сигналов [1], [2]. На первом этапе обработки сигналов использует методы извлечения первичных признак - векторов, которые представляют собой нормированные значения коэффициентов модели линейного предсказания, характеристики основного тона, модели спектра или кепстра. В настоящем исследовании предложено использовать, кепстральные коэффициенты, распределенные по шкале Мэла (MFCC).

Биометрические системы, основанные на цифровой обработке сигналов, не позволяют достичь высокой достоверности идентификации дикторов. Число ошибок растет с увеличением числа дикторов, ухудшением качества сигнала в канале передачи или дополнительными фоновыми шумами. Кроме того, большое значение имеет вариативность голоса человека, которая может привести к изменению голоса до неузнаваемости. Различия в голосе зависит от многих экстралингвистических факторов, таких как степень усталости говорящего, текущее состояние речевого тракта и т.д.

В то же время работы в области биологии показывают способность человеческого мозга практически безошибочно распознавать голос знакомого диктора [18], [19]. Человек имеет самый совершенный на данный момент инструмент идентификации голоса. Поэтому целесообразно разработать техническую систему, моделирующую слуховой аппарат и систему восприятия человека.

Целью исследования является разработка методов, алгоритмов и программ биологически обусловленного текстонезависимого распознавания дикторов, на основе моделирования механизма восприятия речи человеком и использованием класса нейронных сетей с самоорганизацией.

II. Формирование признаков речи

На первом этапе алгоритм выполняет разделение речи и пауз с учетом особенностей восприятия звука человеком [22], которое описывается перcептуальной моделью [23,24]. Перcептуальная модель разделяет весь спектр частот речевого сигнала на отдельные частотные полосы, называемые критическими полосами. Персептуальная модель успешно используется в алгоритмах кодирования акустических сигналов.

В современных компьютерных системах обработки речи для уменьшения шума, связанного с дискретизацией, обычно используют повышенную частоту дискретизации речевого сигнала 22 кГц. При 10%-ной ошибке оценки частотного диапазона речи, начиная от самых низких частот в пределах 50-9500 Гц, его полностью перекрывают 22 полосы персептуальной модели. Таким образом, в системе обработки речевого сигнала следует использовать 22 полосы персептуальной модели.

Кратковременный спектр:



где  - анализируемый сигнал,  - окно,  - кратковременный спектр сигнала.

Кратковременный кепстр сигнала в ой полосе представляет собой обратное преобразование Фурье от логарифма действительной части кратковременного спектра:



Формирование базы высказываний дикторов, достаточной для обучения и тестирования системы распознавания. База должна включать высказывания разных дикторов, в том числе дикторов разного пола, находящихся в разном эмоциональном состоянии, записи могут быть разного качества, содержать шум, иметь разный период дискретизации и шаг квантования. Определение вектора признаков. Сопоставление разных вариантов формирования признаков, полученных из персептуальной модели, таких как кратковременная энергия, коэффициенты линейного предсказания, кратковременные спектральные коэффициенты сигналов в N частотных каналах модели.

1. УПОРЯДОЧЕНИЕ ПРИЗНАКОВ

Мозг человека почти полностью окружен церебральной корой, которая скрывает другие его части. В смысле сложности кора головного мозга, возможно, превышает другую известную структуру. Информация, полученная от сенсоров, отображаются на соответствующую область коры в упорядоченном виде.

Искусственные нейронные сети Кохонена моделируют свойства коры головного мозга упорядоченного представления информации, получаемой от системы восприятия звуков:

* каждый входной образ сохраняется в собственном контексте;
* нейроны, близко расположенные на карте, связаны с близкой звуковой информацией.

Нейронная сеть Кохонена с адаптивными весами получает входной вектор кепстральных коэффициентов  и выполняет отображение этого вектора на множество выходных откликов таким образом, что отклики приобретают топологический порядок. Алгоритм самоорганизации включает инициализацию синоптических весов  малыми случайными значениями, получение последовательности векторов , получение нейрона, выигравшего на текущем шаге  по критерию минимального расстояния до входного образа:



Обновление вектора весов выполняется по следующей формуле:



где - параметр скорости обучения, - функция, характеризующая влияние выигравшего нейрона на его соседей, и динамически изменяющаяся во времени.

В результате упорядочения входные векторы, относящиеся к разным дикторам, образуют отдельные кластеры, которые в общем случае не являются линейно разделимыми. Поэтому для разделения классов может быть использована нейронная сеть одного из следующих видов: многослойный персептрон, нелинейная машина поддержки векторов, нейронные сети с радиальными базовыми функциями (РБФ) [26],[27].

1. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

На основании хранящихся в базе данных образцов семи высказываний двух дикторов выполнено построение карты, показанной на рис.1, на которой красные кластеры соответствуют первому диктору, зеленые – второму. Была использована гексогональная решетка нейронов. Время грубой настройка нейронов составляет 75 минут, время тонкой настройки – 217 минут. Для тестирования были использованы семь других коротких высказываний этих дикторов. В результате тестирования выполнена оценка вероятностей ошибочных решений FQE=1.956, FTE=0.020.

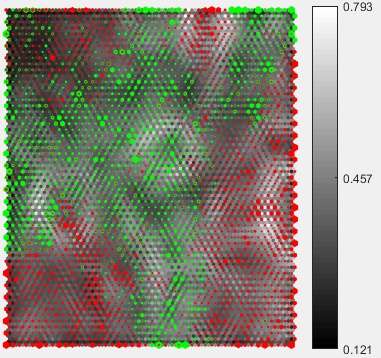


Рисунок 1. Карта коэффициентов кепстра для двух дикторов.

Распознавание дикторов в условиях шума представляет определенные трудности. Проведено исследование влияния белого и розового шума на процесс распознавания. Для одного диктора по семи высказываниям построены карты кепстральных коэффициентов, приведенные на рис.2 и рис.3. Время грубой настройка нейронов составляет соответственно 74 и 72 минут, время тонкой настройки – 219 минут. Тестирование выполнялось по двум высказываниям диктора. Вероятности ошибочных решений FQE=1.904 и 1.863, FTE=0.022 и 0.021.

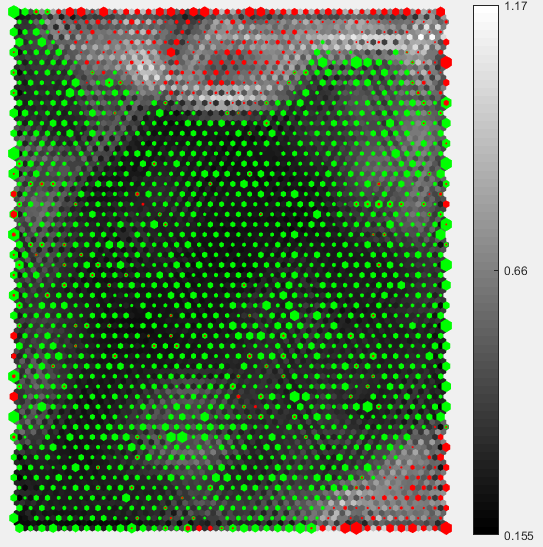


Рис.2. Карта коэффициентов кепстра для одного диктора с добавлением розового шума.

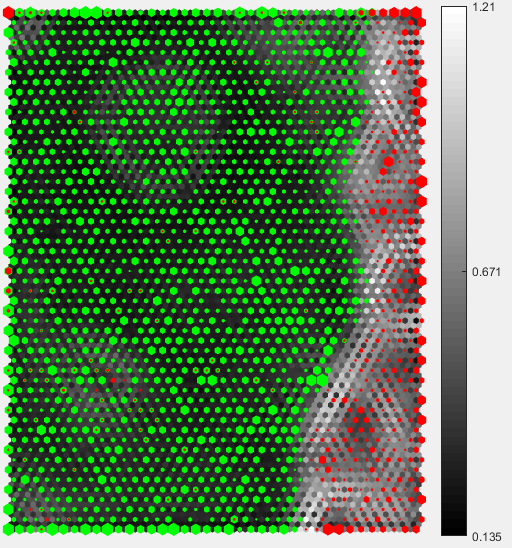


Рис.3. Карта коэффициентов кепстра для одного диктора с добавлением белого шума.

Для анализа больших объемов данных с целью классификации и текстонезависимой идентификации дикторов необходимо использовать высокопроизводительные вычислительные средства. Поэтому вычисления возможно выполнить в суперкомпьютерном центре Санкт-Петербургского политехнического университета Петра Великого.

1. ЗАКЛЮЧЕНИЕ

Предложен новый биологически обоснованный робастный метод распознавания дикторов, моделирующий механизм восприятия речи человеком и использующий биологические принципы в технических системах.

Предложенный метод, в отличие от метода СГР, может работать в реальном времени, поскольку не требует предварительной оценки плотности распределения вероятностей по всем данным эксперимента.

Предложен класс нейронных сетей, моделирующих механизм распознавания дикторов на основе их произвольных высказываний, позволяет повысить достоверность распознавания.

При обработке большого объема данных целесообразно использовать высокопроизводительные вычислительные средства, предоставляемые Суперкомпьютерным Центром «Политехник»

**Публикации**

1. Рабинер Л., Шафер Р. Цифровая обработка речевых сигналов. М.: Радио и Связь, 1981. 492 с.
2. Atal B.S., “Automatic speaker recognition based on pitch contours”, JASA, vol. 52, pp. 1687-1697, 1972
3. Carey M.J., Parris E.S., Lloyd-Thomas H., and Bennet S., “Robust prosodic features for speaker identification”, Proc. ICSLP-96, Philadelphia, Nov. 1996
4. Sonmez K., Shriberg E., Heck L., and Weintraub M., “Modeling dynamic prosodic variation for speaker verification”, Proc. ICSLP-98, Sydney, Dec 1998
5. Weber F., Manganaro L., Peskin B., and Shriberg E., “Using prosodic and lexical information for speaker identification”, Proc. ICASSP-02, Orlando, May 2002
6. Andrews W., Kohler M., Campbell J., Godfrey J., Hemandez-Cordero J., “Gender dependent phonetic refraction for speaker recognition”, ICASSP, vol. 1, p. 149-152, 2002
7. Navratil J., Jin Q., Andrews W., Campbell J., “Phonetic speaker recognition using maximum likelihood binary decision trees”, ICASSP 2003
8. Arcienega M., Alexander A., Zimmermann P., Drygajlo A. Bayesian Network Approach Combining Pitch and Spectral Envelope Features to Reduce Channel Mismatch in Speaker Verification and Forensic Speaker Recognition // Proceedings of Interspeech'2005, Lisbon, Portugal, Sept. 4-8, 2005. P. 2009-2012
9. Rose 1990] Rose, R.C. and Reynolds, D.A., Text-independent speaker identification using automatic acoustic segmentation. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1990, pp. 293–296.
10. Reynolds, D.A. and Rose, R.C., Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1995), pp. 72–83.
11. Reynolds, D.A., Automatic speaker recognition using Gaussian mixture speaker models, Lincoln Lab. J. 8 (1996), pp. 173–192.
12. TeuvoKohonen, Self-Organized Formation of Topologically Correct Feature MapsBioljgical Cybernetics. 43, 1982, pp. 59-69.
13. Haykin, S., 2009. Neural Networks and Learning Machines. 3rd edition ed. Pearson Education, Inc.
14. Yakovenko, A., Malykhina, G.F.Text-independent speaker recognition using radial basis function network Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9719,2016 , pp.74-81.
15. A.A. Yakovenko, G.F. Malykhina. Radial-Basis Functions Neural Network for Text Independent Speaker Identification. Proceedings of the 2014 International Conference on Methods in Applied SAciences (MMAS’14), pp 389-392.