Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 8, Пенза-2012, с. 73-75 (http://пниэи.рф/activity/science/BIT/T8-p73.pdf)

РЕЧЕВОЙ ФРАГМЕНТАТОР ДЛЯ НЕЙРОСЕТЕВОГО БИОМЕТРИЧЕСКОГО ВОКОДЕРА

Гришин В.М., Калашников Д.М. (Пенза)

Для осуществления операций идентификации диктора по произносимой парольной/произвольной фразе необходимо в первую очередь выполнить преобразование речи человека в цифровой поток. Этой операцией занимаются вокодеры, имеющие различные варианты реализации.

В речевых элементных вокодерах при кодировании распознаются произносимые элементы речи (например, фонема) и на выход кодера подаются только их номера [1]. В таких вокодерах происходит автоматическое распознавание слуховых образов, а не определение параметров речи и, соответственно, теряются все индивидуальные особенности говорящего.

В контексте биометрической идентификации диктора по произносимой парольной/произвольной фразе необходимо повышать узнаваемость речи говорящего за счет учета дополнительных параметров, соответственно, увеличивая размерность решаемой задачи.

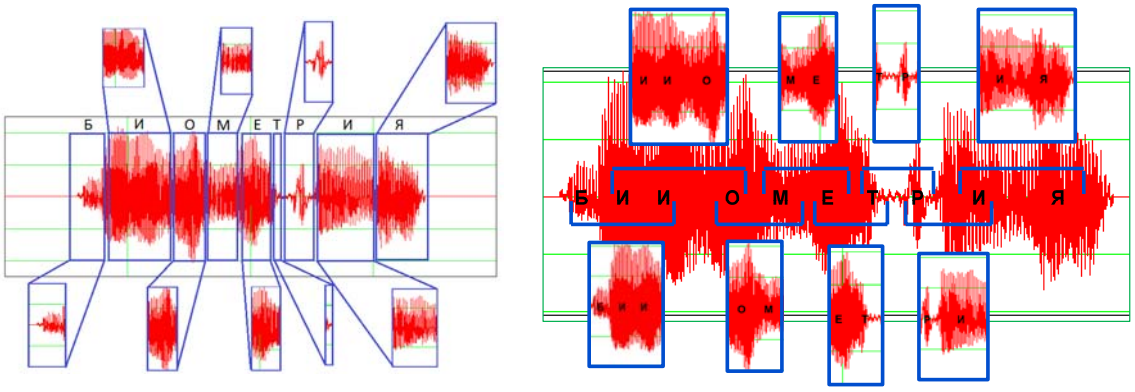
Современные вокодеры, очевидно, не способны решать высокоразмерные задачи. В настоящее время, когда сняты ограничения по вычислительным мощностям, свои ограничения накладывает линейная алгебра, заложенная в вокодерах. В связи с этим стоит отказаться от вокодеров старого образца, которые ввиду ужесточения требований к системам идентификации диктора по голосу, вскоре перестанут справляться с поставленными перед ними задачами, и перейти к реализации вокодеров нового поколения — биометрических вокодеров, в которых будут сняты ограничения на размерность решаемых задач, потому, что в данный момент появилась возможность обучать нейронные сети.

Применяя биометрические вокодеры на нейросетевых функционалах можно будет увеличивать размерность решаемой задачи до тех пор, пока результат не будет удовлетворять поставленным целям.

Процесс обучения каждого искусственного нейрона должен обеспечивать высокий уровень вероятности принятия биометрического образа «Свой» при низком уровне вероятности ошибочного принятия биометрического образа «Чужой». Следовательно, должна быть сформирована база биометрических образов для обучения. Необходимо провести оценку размера данной базы. Также следует сразу принять во внимание, что наряду с огласованными фонемами в ней должны содержаться примеры всех шипящих звуков, так как недавние исследования в этой области показали, что одинаковые шипящие звуки, воспроизводимые разными людьми, обладают существенными спектральными отличиями [2].

Исходя из вышеизложенного, на первый план выходит проблема реализации некоего речевого фрагментатора, в задачу которого входит организация процесса разбиения непрерывного речевого сигнала на отдельные звуки либо фонемы, для последующего формирования из них базы биометрических образов, применяемых в обучении биометрического вокодера. На рисунке 1 представлен пример различных вариантов фрагментации заданного парольного слова. Все описанные ниже эксперименты проводились с привлечением двух дикторов, мужского и женского пола, произносящих

конкретное парольное слово. Данные получены при частоте дискретизации 8000 Гц 8-ми разрядным АЦП.



а) фрагментация по отдельным звукам

б) фрагментация по фонемам

Рисунок 1 – Варианты фрагментации заданного парольного слова.

Практика показала, что аутентификация оказывается достаточно надежной, если длина фиксированной голосовой фразы составляет три слова длинной от 4 до 9 букв, при том, что для обучения автомата потребуется предъявить порядка 20 примеров заданной голосовой фразы. При средней длине слова 7 букв получается, что для успешного обучения искусственного подсознания биометрического вокодера требуется произнесения диктором «Свой» текста примерно из 420 букв (звуков) или примерно 10 строк текста.

Если идентифицировать диктора необходимо на произвольном тексте, то придется использовать пары наиболее часто встречающихся в языке звуков. По данным, приведенным в работе [3], вероятность наиболее часто встречающихся пар букв русскоязычного текса составляет величину близкую к 0,01. То есть, для получения от 10 до 20 примеров часто встречающихся пар звуков (букв), потребуется использовать произвольный обучающий текст длиной от 1000 до 2000 букв (звуков). При темпе речи 10 звуко-букв в секунду необходимо будет иметь образец речи диктора длинной от 100 до 200 секунд (от 1,7 до 3,3 - минут).

В первую очередь, системе необходимо определить, сколько звуков содержится в поступившем входном сигнале парольной/произвольной фразы. Для этого нужно вычислить огибающую амплитуды речевого сигнала. В результате получится функция, график которой описывает закон изменения во времени амплитуды исходного сигнала. Соответственно, считая «скачки» экстремумов, система определит примерное количество звуков, содержащихся в сигнале.

Очевидно, что в данной системе будет использоваться покадровая обработка речевого сигнала. Но кадры должны быть плавающие, кратные периоду основного тона выделяемого звука.

Характеристики голосового тракта можно считать неизменными на интервале 10-20 мс, то есть параметры надо измерять с частотой порядка 50 Гц.

Очевидно, чтобы определить момент начала следующего звука в речевом сигнале, фрагментатор должен отслеживать изменение периода основного тона в сигнале. Здесь следует отметить, что как правило величина периода основного тона человека незначительно, но колеблется во времени. Его изменения обычно укладываются в интервал ±20%. Следовательно, данный интервал должен учитывать и речевой фрагментатор, чтобы не спутать начало следующего звука в речевом сигнале с обычным незначительным изменением периода основного тона текущего звука. Это же должно учитываться и при выборке, сравнении и записи речевых фрагментов в базу образов.

Важно не только реализовать синхронизацию каждого кадра фрагментатора с началом звука, но и с серединой звука. Очевидно, что именно в середине звука, особенно огласованного, присутствуют гармоники, наиболее полно описывающие именно индивидуальные особенности речи диктора. Система должна провести анализ звука и, если он огласован, синхронизировать кадр фрагментатора с отсчетом, имеющим наибольшую амплитуду, потому что в большинстве случаев именно в середине звука присутствуют отсчеты с максимальной амплитудой.

Если же звук взрывной («т», «п» и т.д.), то очевидно, что образец необходимо брать вначале звучания, так как, исходя из рисунка 2, именно здесь у подобных звуков присутствуют гармоники, наиболее полно описывающие именно индивидуальные особенности речи диктора.

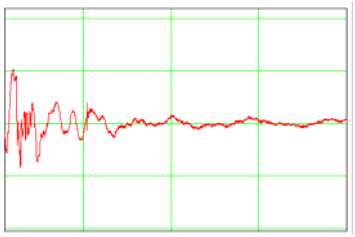


Рисунок 2 – Амплитуда звука «т»

Соответственно, во фрагментаторе должен быть реализован механизм, выделяющий из звукового сигнала «взрывные» звуки.

Стоит отметить, что на данном этапе не идет речь о распознавании звуков. Речь идет о примитивном делении непрерывного входного речевого сигнала на фрагменты. Потому, что без данных фрагментов невозможно будет реализовать последующее обучение нейросети.

Особенностью и достоинством данного фрагментатора является то, что это низкоуровневый механизм, который будет моментально обучаться сам на себе, просто сравнивая текущий звук с соседями и параллельно с этим создавая базу биометрических образов для их дальнейшего использования в обучении нейросетевых биометрических вокодеров.

Список использованной литературы:

1. Шелухин О.И., Лукьянцев Н.Ф. «Цифровая обработка и передача речи». М., «Радио и связь», 2000.

2. Иванов А.И., Хальметова А.Н., Захаров О.С., Рыболовлев А.А., Рыжаков А.П. Нейросетевой вокодер-архиватор, сохраняющий биометрические особенности голоса говорящего при высоком уровне сжатия шипящих звуков //Нейрокомпьютеры, разработка, применение. – 2012. – №3. – С. 44-49

3. Елфимов А.В., Воячек С.А., Качайкин Е.И., Куликов С.В. Обучение нейросетевого идентификатора авторства рукописных текстов. «Нейрокомпьютеры: разработка, применение» №6, 2009 с. 17-21

Статья поступила 20.12.2012, опубликовано в Интернет 15.01.2013 по положительной рецензии д.т.н., доцента Иванова А.И.