электронный научно-технический журнал
ИНЖЕНЕРНЫЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51036. ISSN 2307-0595

Анализ подходов к решению задачи распознавания интенсивных кратковременных звуков

 # 03, март 2015

профессор Иванова Г. С.[1,*], Кожушко В. В.[1]

УДК: 004.052

[1]Россия, МГТУ им. Н.Э. Баумана
[*]gsivanova@gmail.com

Введение

Задача распознавания звуков различного происхождения встречается в следующих вариантах:

1) распознавание человеческой речи,
2) фильтрация помех и выделение человеческой речи,
3) распознавание звуков не речевого происхождения.

Частным случаем распознавания звуков не речевого происхождения является задача распознавания интенсивных кратковременных звуков (ИКЗ). Под ИКЗ будем понимать звуковые колебания длительностью менее 0,5 секунды, громкость которых значительно выше (на 15 и более дБ), чем фоновые звуки. Примерами таких звуков являются хлопки, выстрелы, взрывы.

Программам, распознающим ИКЗ, можно найти широкое применение, например:

- в промышленности – определение хлопков, взрывов для мгновенного запуска системы защиты в случае аварии или повреждении емкостей с химически опасными веществами,
- в космонавтике – для своевременной реакции операторов на изменения в работе оборудования,
- на испытательных полигонах или в стрелковых тирах – для автоматизации работы с контрольно-измерительным оборудованием,
- в охранных системах – для подачи сигналов о нарушении защиты,
- в системах «Умный дом» – как для распознавания управляющих воздействий, так и с целью получения информации о событиях, требующих внимания, например распахнувшемся от ветра окне.

Анализ литературы по данному вопросу показал, что больше всего достижений имеется в области распознавания человеческой речи. При решении этой задачи требуется вос-

становлении по звуковому сигналу слова естественного языка из ограниченного словаря, произнесением которого является этот звуковой сигнал. Она решается путем создания шаблонов эталонных слов (словаря) и последующим сравнением с ними звуковых сигналов [1].

Методика распознавания речи предполагает, что сначала запись сигнала разбивают на части одинаковой длины. Полученные фрагменты записи преобразуют из временной области в частотную (например, с помощью преобразований Фурье) [2-5], чтобы близость отрезков относительно простых метрик соответствовала близости участков сигналов «на слух». Следующий этап – нахождение соответствия между промежутками звукового сигнала и окнами эталонных слов. Сложность заключается в том, что различные участки звукового сигнала при воспроизведении одного и того же слова отличаются разной скоростью произношения. Дополнительные трудности вносят особенности человеческого голоса, который может быть уникальным, имеет особенный акцент, а так же имеет собственную частоту колебания голосовых связок.

Фонетические модели, используемые в программировании не точны, так как не учитывают всего многообразия факторов. Для задания фонетических эталонов обычно используют статистические методы, предполагающие, что акустические параметры фонем распределены по нормальному закону [1,6]. В действительности картина намного сложнее: точная модель эталонов звуков и слов должна включать в себя множество элементов (по одному на каждый вариант произнесения).

В отличие от задачи распознавания человеческой речи, которая, прежде всего, ориентирована на дикторонезависимость, задача распознавания ИКЗ обладает следующими особенностями:

- анализируемый диапазон частот существенно меньше,
- нижняя граница значения интенсивности громкости звука имеет более высокое значение, чем в речевых сигналах,
- длительность ИКЗ составляет менее половины секунды.

Таким образом, алгоритмы распознавания речи, предназначенные для распознавания длительных звуков сложной конфигурации, для распознавания ИКЗ могут оказаться не применимыми или неэффективными. Следовательно, необходимо разработать алгоритмы, позволяющие эффективно распознавать ИКЗ.

1 Постановка задачи и способы представления звука

В зависимости от конкретного варианта применения могут быть востребованы следующие функции подсистемы распознавания ИКЗ:

- измерение временных интервалов между определенными ИКЗ,
- выделение конкретного источника звука из звукового эфира,
- распознавание ИКЗ в условии шумов.

Наибольший интерес представляют подзадачи выделения конкретного источника звука и распознавания заданных звуков в условии шумов. Если человек может отфильтро-

вать весь «мусор» и помехи в звуковом эфире, настраиваясь на определенные тоны, голоса, музыку, то в случае машинного распознавания задача выделения источника звука требует специальных методов решения.

Выбор метода решения перечисленных выше задач может существенно отличаться в зависимости от используемого формата представления звуковых данных.

1.1 Представление звука

Цифровым представлением звука является массив чисел, содержащий отсчёты или выборки, которые соответствуют величине звукового давления в последовательные моменты времени (рисунок 1).

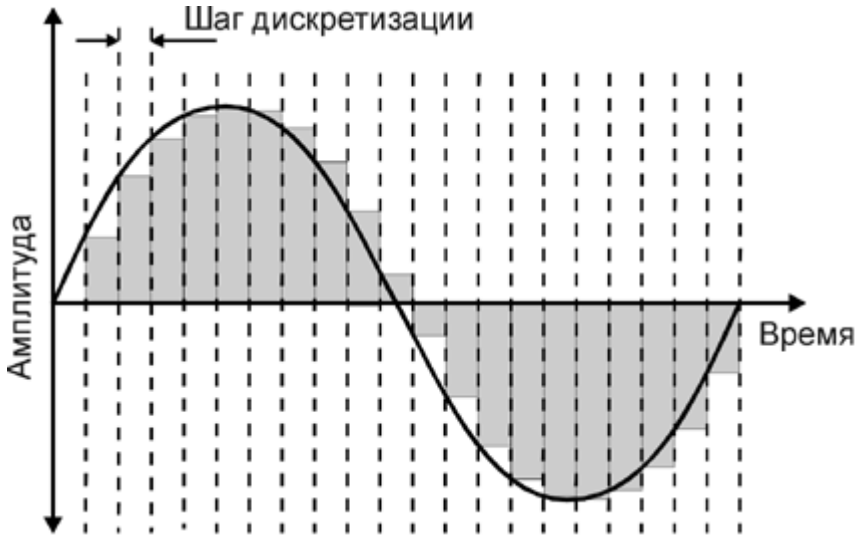


Рисунок 1 - Цифровое представление звука

Значения в цифровом массиве соответствуют определенному промежутку времени – шагу дискретизации. Следует также учитывать, что для записи высоты каждого блока используют дискретный набор значений. Соответственно, высоты блоков могут не совсем точно совпадать с волной, что может привести к погрешностям [2].

Таким образом на практике звук представлен набором значений звукового давления во временной области с выбранным шагом дискретизации (рисунок 2). Конкретная форма записи и хранения указанных наборов значений зависит от используемого аудио формата.

1.2 Выбор аудио формата для анализа звука

При решении перечисленных выше задач распознавание звука должно происходить в режиме реального времени. Это предполагает многократное повторение трех операций: записи порции звуковых колебаний, обработки полученных значений и вывода результата. Для снижения времени анализа звуковых данных необходим аудио формат, который позволит максимально быстро обрабатывать данные.

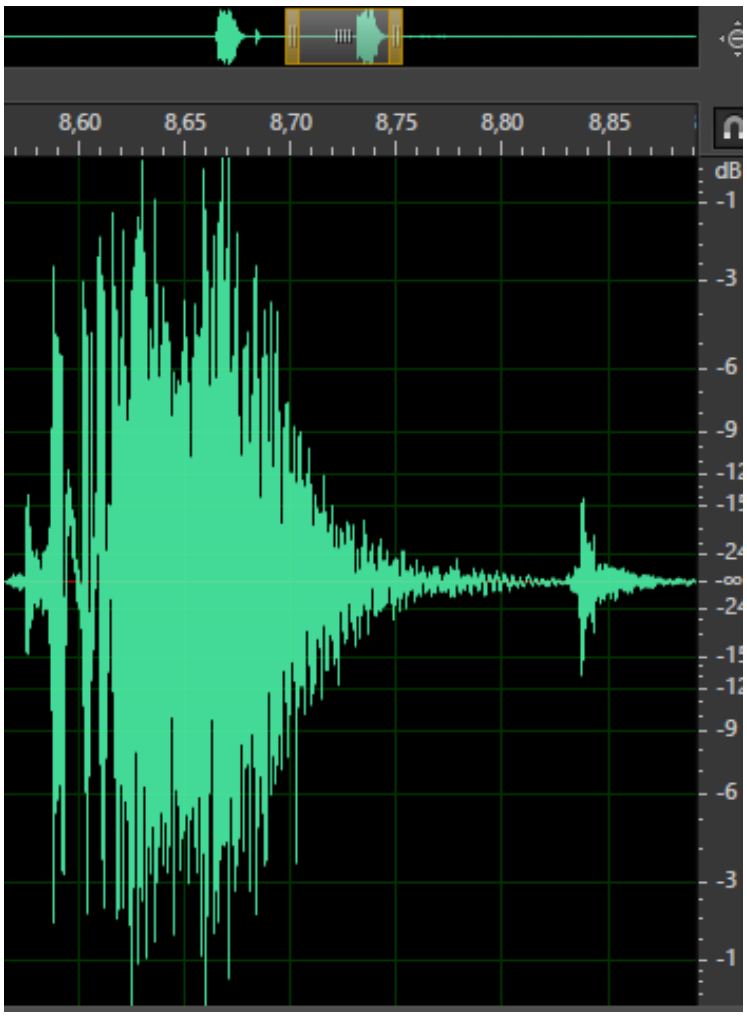


Рисунок 2 – Представление звука выстрела во временной области в формате mp3

Формат представления звуковых данных в цифровом виде зависит от способа квантования аналогово-цифровым преобразователем.

Выделяются следующие группы аудио форматов [7]:

- аудио форматы без сжатия (*wav*, *aiff*);
- аудио форматы со сжатием без потерь (*ape*, *flac*);
- аудио форматы, с применением сжатия с потерями (*mp3*, *ogg*).

Форматы, использующие сжатие, предназначены, прежде всего, для хранения и воспроизведения звуковых данных, предназначенных для прослушивания человеком, слух которого весьма не совершенен.

Для высокоскоростного и сравнительно точно анализа форматы со сжатием, и уж тем более с потерями непригодны. Наличие потерь неприемлемо, так как снижает точность получаемых результатов, а необходимость распаковки данных – увеличивает время их обработки. Следовательно, необходимо использовать формат без компрессии.

Анализ показал, что всю необходимую информацию в удобном для обработки виде содержит формат *WAVE* (*wav*).

Аудиофайл *wave* состоит из двух блоков. Первый блок — информационный заголовок файла, в котором содержится следующая информация:

- размер файла,
- количество каналов,
- частота дискретизации,
- глубина звучания (битрейт).

Второй блок состоит из данных, хранящих цифровой сигнал – набор значений амплитуд. Структура wave файла представлена в таблице 1.

Таблица 1 – Структура *wave* файла

Номер байта	**Поле**	**Описание**
0..3 (4 байта)	*chunkId*	Содержит символы "*RIFF*" в ASCII кодировке.
4..7 (4 байта)	*chunkSize*	Размер файла
8..11 (4 байта)	*format*	Содержит символы "*WAVE*"
12..15 (4 байта)	*subchunk1Id*	Содержит символы "fmt"
16..19 (4 байта)	***subchunk1Size***	16 для формата для импульсно-кодовой модуляции.
20..21 (2 байта)	***audioFormat***	Аудио формат.
22..23 (2 байта)	***numChannels***	Количество каналов.
24..27 (4 байта)	***sampleRate***	Частота дискретизации.
28..31 (4 байта)	***byteRate***	Количество байт, переданных за секунду воспроизведения.
32..33 (2 байта)	***blockAlign***	Количество байт для одного сэмпла, включая все каналы.
34..35 (2 байта)	***bitsPerSample***	Глубина звучания
36..39 (4 байта)	***subchunk2Id***	Содержит символы "*data*"
40..43 (4 байта)	***subchunk2Size***	Количество байт в области данных.
44..	***Data***	*WAV*-данные.

Как следует из таблицы, аудио формат *wav* позволяет записывать сигнал сразу по нескольким каналам и с разной частотой дискретизации. Человеческий слух воспринимает звуки с частотой дискретизации не более 22 кГц, поэтому рабочая частота должна находиться в тех же пределах. Кроме того не целесообразно ориентировать подсистему на обработку сигналов нескольких каналов, поскольку в большинстве вычислительных устройств со встроенными возможностями звукозаписи применяются одноканальные микрофоны со средним качеством записи, а использование нескольких каналов увеличит время обработки. Таким образом, определены следующие параметры аудио файла:

- одноканальный режим записи (моно),
- 16 битное квантование,
- частота дискретизации 22050 Гц.

Помимо решения вопроса, связанного с представлением информации, необходимо оценить влияние условий среды, в которой распространяется звук, на точность распознавания ИКЗ.

1.3 Факторы, влияющие на корректность распознавания ИКЗ

Поскольку распознавание звука происходит в воздушной среде, точность распознавания зависит от:

1) температуры воздуха,
2) влажности воздуха,
3) атмосферного давления,
4) расстояния между датчиком и источником звука,
5) геометрических параметров пространства (помещения или улицы).

Оценим влияние указанных факторов на корректность распознавания.

Экспериментальные оценки зависимости уровня громкости звука (УГЗ) от *расстояния* между источником звука и датчиком в узком длинном помещении представлены в таблице 2, а от *температуры* воздуха в помещении – в таблице 3. Для повышения точности измерений значения фиксировались для пяти ИКЗ (выстрелы) и затем усреднялись.

Таблица 2 – Зависимость УГЗ от расстояния между источником звука и датчиком

Расстояние между источником звука и датчиком, м	**Уровень звука, дБ**					**Среднее значение уровня звука, дБ**	**Снижение, %**
	Номер эксперимента						
	1	**2**	**3**	**4**	**5**		
0,5	84	84	84	82	86	84	0,00
1	79	80	79	80	82	80	4,76
1,5	79	81	79	79	79	79,4	0,75
2	79	79	80	79	79	79,2	0,25
2,5	79	79	79	79	79	79	0,25
3	79	79	78	79	79	78,8	0,25
3,5	78	79	78	79	79	78,6	0,25
4	78	78	79	79	78	78,4	0,25
4,5	79	78	79	78	78	78,4	0,00
5	78	78	79	78	77	78	0,51

Таблица 3 – Зависимость УГЗ от температуры

Температура t, С	Уровень звука, дБ					Среднее значение уровня звука, дБ	Снижение, %
	Номер эксперимента						
	1	2	3	4	5		
26	83	82	84	82	83	82,8	
24	83	82	82	83	83	82,6	0,24
22	83	82	82	82	83	82,6	0,00
20	82	83	82	82	83	82,4	0,24
18	83	82	81	82	83	82,2	0,24

Зависимость УГЗ от расстояния между датчиком и источником звука представлена на рисунке 3.

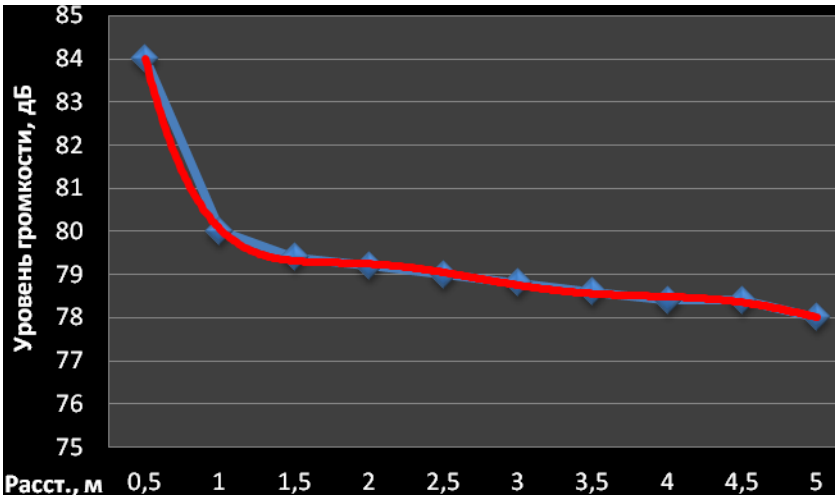


Рисунок 3 –Зависимость УГЗ от расстояния между датчиком и источником звука, полученная экспериментально

Анализ полученных данных показал, что снижение УГЗ на расстоянии более чем 1,5 метра между датчиком и источником не значительно, в то время, когда на дистанциях менее 1,5 метра наблюдается существенное снижение уровня громкости звука при удалении от датчика.

Теоретически разность уровней громкости между двумя точками, удаленными на разное расстояние от источника звука, описывается формулой [6]:

$$L_1 - L_2 = 20 \times \log \frac{R2}{R1}, \quad (1)$$

где L_1 и L_2 – уровни громкости, дБ, R_1 и R_2 – расстояния между источником звука и приемником (рисунок 4).

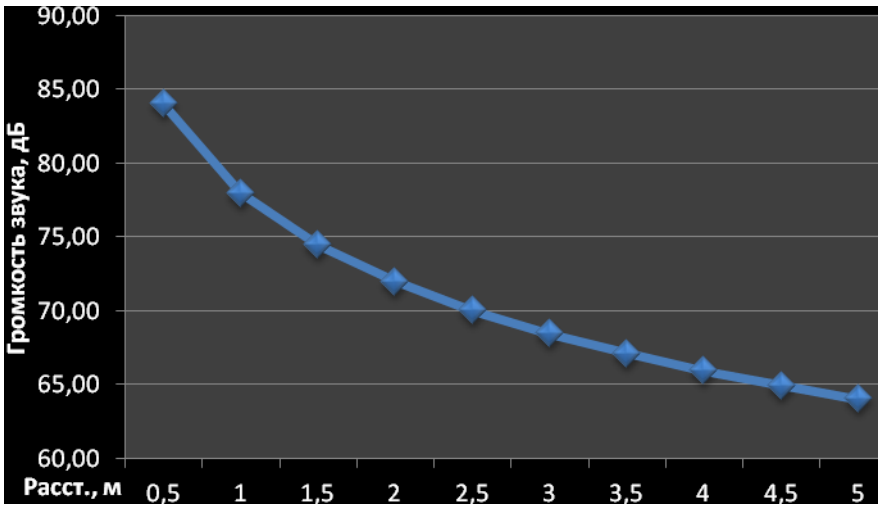


Рисунок 4 - Зависимость УГЗ от расстояния между датчиком и источником звука, рассчитанная теоретически

Несовпадение практически полученных результатов и теоретических расчетов связано с тем, что при измерении на конечный результат повлияли геометрические параметры помещения. При проведении эксперимента эффект реверберации усилил громкость звука на коротких расстояниях [8].

На рисунке 5 показана зависимость УГЗ от температуры воздуха. Изменение УГЗ при снижении температуры по сравнению со снижением УГЗ при удалении от источника не существенно, соответственно им можно пренебречь.

Влажность и давление воздуха также можно не учитывать, поскольку они не оказывают несущественного влияния на УГЗ [10].

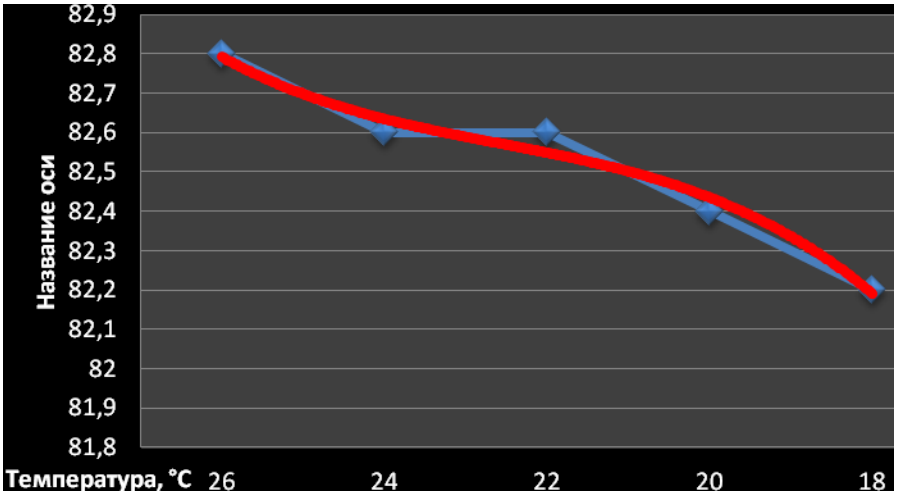


Рисунок 5 – График зависимости УГЗ от температуры воздуха

Таким образом, согласно полученным экспериментальным оценкам, при разработке подсистемы распознавания ИКЗ необходимо учитывать:

1) геометрические параметры помещения,
2) расстояние между источником звука и датчиком.

2 Выбор метода распознавания ИКЗ

2.1 Анализ применимости методов распознавания речи для распознавания ИКЗ

Для решения задачи распознавания ИКЗ могут быть использованы следующие методы распознавания речи:

1) частотный метод, основанный на применении частотного анализа;
2) метод, основанный на применении нейронных сетей.

Рассмотрим, как перечисленные методы можно применить для распознавания ИКЗ на примере выявления звуков выстрелов.

На рисунке 6 представлен график зависимости уровня громкости звука от времени при выполнении 5 выстрелов. В качестве приемника звука при получении этой записи использовался смартфон *Samsung GT*-I9100 (*Galaxy S*-II) с приложением *Smart Tools* версии 1.7.7, а в качестве источника ИКЗ – пневматический пистолет МР 654k 2006 года выпуска заводской комплектации.

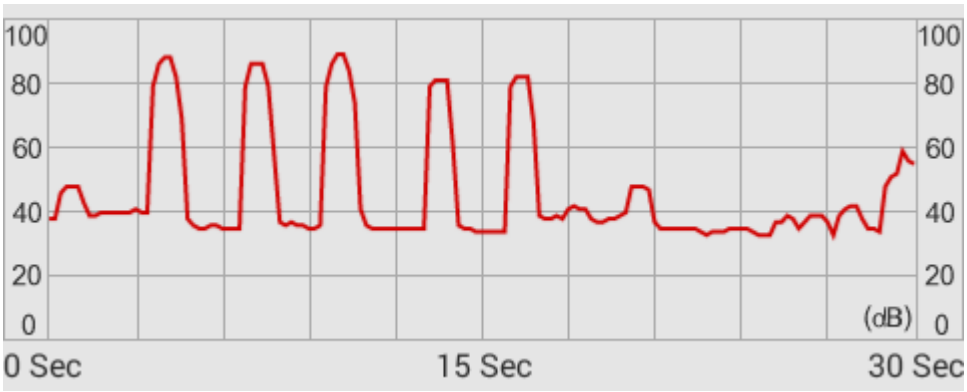


Рисунок 6 – Зависимость величины УГЗ от времени при выполнении 5 выстрелов

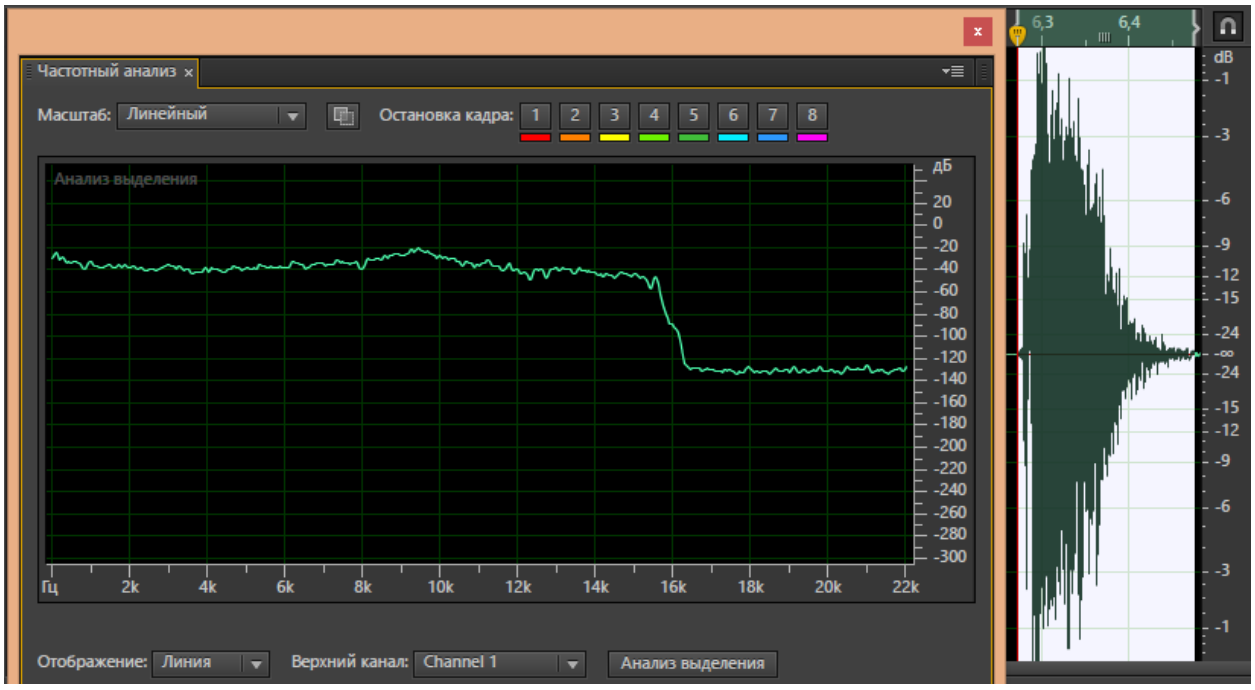
Анализ полученных данных показал что записи звуков выстрелов различаются по длительности и амплитуде примерно на 15-20 %, и при этом амплитуда ИКЗ примерно в два раза превышает амплитуды шумовых сигналов.

2.1.1 Частотный метод

Частотный анализ предполагает переход из временной (см. рисунок 1) формы представления звука в частотную, где каждой величине частоты соответствует некоторая величина звукового давления. На рисунке 7 показаны частотные представления ИКЗ вы-

стрелов пневматических пистолетов Аникс А101 и МР 654k, полученные с помощью смартфона *Xiaomi Red Rice 1S* и программы *Adobe Audition CC*6.

а



б

Рисунок 8 – Частотное представление звука выстрела для пневматических пистолетов: *а* – Аникс А101; *б* – МР 654k

В частотной области для каждого звука значения звукового давления уникальные, что позволяет с высокой эффективностью применить его для задачи распознавания ИКЗ.

Критерием для сравнения является набор значений звукового давления на заданных интервалах частот.

Однако использование этого метода предполагает дополнительный этап вычислений, что может быть не применимо для анализа звука в режиме реального времени.

2.1.2 Распознавание ИКЗ с помощью нейронных сетей

При распознавании звуков с использованием нейронных сетей решается задача распознавания образов по шаблону[11,12]. При этом каждый нейрон (рисунок 9) представляет собой эталонное значение зависимости звукового давления от времени [13, 14].

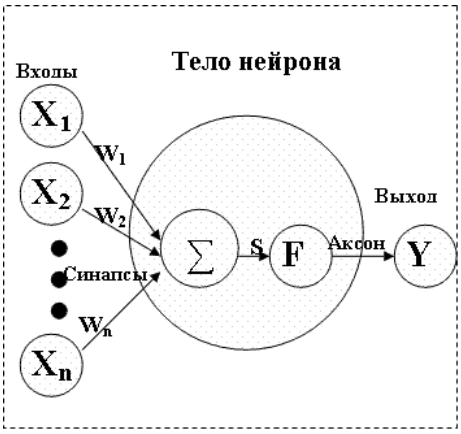


Рисунок 9 - Представление нейрона

Синоптические связи характеризуются весами w_i, на вход которых подается значение УГЗ на некотором интервале. Текущее состояние S нейрона равно взвешенной сумме входов:

$$S = \sum_{i=1}^{n} x_i w_i \, .$$

Функция S далее преобразуется активационной функцией F и дает выходной сигнал Y нейрона. Самой распространенной в нейронных сетях активационной функцией является логистическая функция, вычисляемая как:

$$F(S) = \frac{1}{1+e^{-s}} \, .$$

Главным преимуществом нейронных сетей является возможность их предварительного обучения, что существенно увеличивает вероятность правильного распознавания.

С точки зрения распознавания ИКЗ необходимость обучения нейронной сети и, следовательно, создание базы шаблонов – недостаток, поскольку без начальных данных нет возможности быстро ввести оборудование в эксплуатацию.

Издержки применения методов распознавания речи для распознавания ИКЗ заставляют искать другие способы решения этой задачи.

2.2 Метод фильтрации уровней громкости ИКЗ

Форма сигналов ИКЗ существенно проще, чем форма речевых сигналов, поэтому для распознавания ИКЗ можно предложить метод, основанный на разнице уровней громкости ИКЗ и фона.

Для выделения ИКЗ задается фиксированное значение громкости звука L и смещение ΔL в процентах от L (рисунок 10). Полученные значения $L+\Delta L$ и $L-\Delta L$ являются верхней и нижней границей чувствительности микрофона соответственно. Таким образом, требуемые звуки попадают в диапазон и распознаются, а шум срезается.

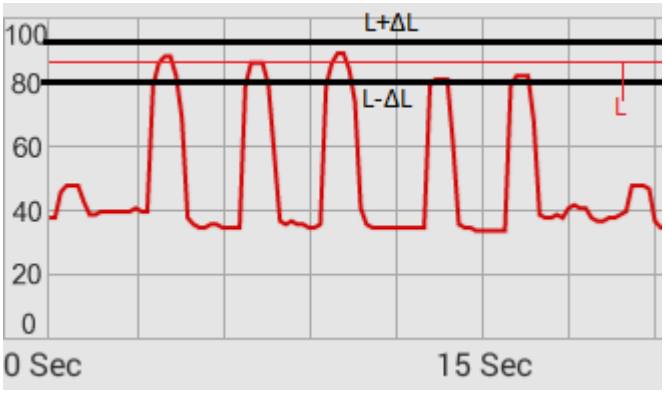


Рисунок 10 – Фильтрация уровней громкости

Смещение ΔL относительно граничного значения, должно определяться экспериментально в зависимости от природы звука.

Способ очень прост в реализации, однако для высокой надежности распознавания требует верного определения смещения ΔL, поскольку даже у одного и того же источника звука уровень громкости может существенно колебаться (см. пики 3 и 4 на рисунке 6). Датчик может реагировать на другие источники звука, которые имеют схожие характеристики, либо отфильтровать необходимый звук, поскольку он не попал в заданный интервал. Но, с другой стороны, можно предположить, что при правильном определении смещения, а так же при учете полученной зависимости громкости звука от расстояния, данный способ будет давать высокую эффективность распознавания в условии шумов или большого количества источников звука со схожей природой на коротких дистанциях.

2.3 Направленное прослушивание эфира

Повысить надежность распознавания ИКЗ можно, используя направленное прослушивание эфира двумя микрофонами. При этом пространство делится на n равных частей, например 8. Используя данные, полученные с двух микрофонов устройства, определяется сектор, в котором находится источник звука. Вещание с остальных направлений игнорируется.

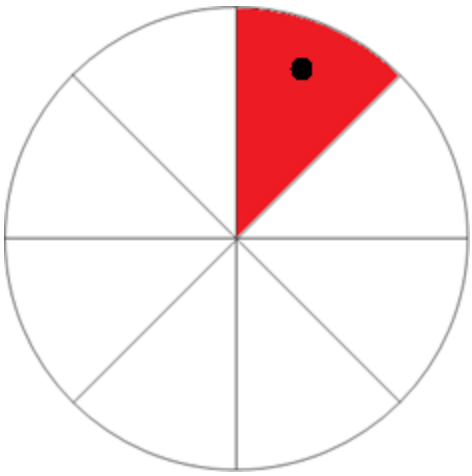


Рисунок 7 – Определение направления на источник звука

Этот метод может не давать высокой точности, поскольку результаты зависят от параметров используемых устройств, которые определяют:

- качество звукозаписи,
- программную доступность двух микрофонов.

В зависимости от добротности записи звука, полученной с микрофона, может неверно определяться положение источника звука в пространстве (попадание в другой сектор). Дополнительно, возможен прием вещания от других объектов, находящихся в той области пространства, в которой расположен необходимый источник. Так же, не всегда имеется возможность программной (для поставленной задачи) реализации, поскольку не каждое устройство имеет 2 программно доступных микрофона.

2.4 Сравнительный анализ методов распознавания и особенности их реализации для решения задачи распознавания ИКЗ

Выбор способа распознавания определяется в зависимости от требований к быстродействию, эффективности, времени на разработку. Таблица 3 содержит экспертную оценку применимости рассмотренных методов для распознавания ИКЗ.

Таблица 3. Применимость способов для распознавания ИКЗ

Способ	Эффективность	Сложность реализации	Быстродействие
Частотный анализ	высокая	средняя	среднее*
Нейронные сети	высокая**	высокая	среднее
Фильтрация уровней громкости звука	средняя	низкая	высокое
Направленное прослушивание	средняя	средняя	высокое

* – требуется дополнительный этап вычислений;

** – после предварительного обучения.

Таким образом, наиболее надежно ИКЗ можно распознать, используя нейронные сети. Однако при наличии ограничений по времени распознавания предпочтительно использовать направленное прослушивание эфира и метод фильтрации уровней громкости звука.

При программной реализации методов распознавания следует учитывать, что запись звука в режиме реального времени производится выборками с заданной частотой. Поэтому алгоритм распознавания должен выполнять поиск ИКЗ на звуковой дорожке с учетом возможного разделение звукового сигнала границей выборки.

Для поиска требуемых звуков можно предложить разбивать данные звуковой дорожки на интервалы, затем анализировать каждый полученный интервал на наличие требуемых звуков. Для исключения потери помимо проверки выборок необходимо проверять интервал, полученный соединением двух соседних выборок.

Заключение

Анализ и экспериментальные измерения, приведенные в статье, позволяют сделать следующие выводы:

- распознавание ИКЗ имеет ряд особенностей, которые выделяют его в отдельную от распознавания речи задачу, актуальность решения которой не вызывает сомнения;

- запись ИКЗ для последующего распознавания в режиме реального времени целесообразно выполнять в формате wave, который не предполагает потерь и сжатия;

- наибольшее искажение при распространении ИКЗ связано с геометрической формой помещения и расстоянием до источника звука, соответственно эти характеристики следует учитывать при распознавании, в то время как температурой воздуха, его влажностью и давлением можно пренебречь;

- сравнительно простая форма звукового сигнала ИКЗ позволяет выполнять распознавание методом фильтрации уровней громкости, для увеличения точности результатов которого возможно использование направленного прослушивания эфира, однако при увеличении требований к надежности распознавания следует использовать методы распознавания речи, т.е. нейронные сети.

Список литературы

1. Фролов А.В., Фролов Г.В. Синтез и распознавание речи. Современные решения.– Режим доступа: http://www.frolov-lib.ru/books/hi/ch05.html (дата обращения: 10.01.2015).
2. Кинтцель Т. Руководство программиста по работе со звуком = A Programmer's Guide to Sound: Пер. с англ. – М.: ДМК Пресс, 2000. 432 с, ил. (Серия «Для программистов»).
3. Распознавание речи. Режим доступа: http://habrahabr.ru/post/226143/ (дата обращения 1.12.2014 г.).
4. Костромицкий С.П. Распознавание длящихся звуков речи // Вестник Тамбовского университета. Серия Естественные и технические науки. 2003. №1. С. 204.

5. Компьютерное распознавание и порождение речи. Режим доступа: http://speech-text.narod.ru/chap4_1_1.html (дата обращения: 10.01.2015).
6. Савельев И.В. Курс общей физики. Механика, колебания и волны, молекулярная физика. Том I М.: Наука, гл. ред. физ-мат. лит., 1970. — 508с.
7. Сравнение цифровых аудиоформатов. Режим доступа: https://ru.wikipedia.org/wiki/ (дата обращения: 10.01.2015).
8. Реверберация звука. Режим доступа: http://wikisound.org/Реверберация (дата обращения 10.01.2015).
9. Elinek F. Распознавание непрерывной речи статистическими методами // ТИИЭР 64. 1976. №4. С.131-160.
10. Кульков Я.Ю., Кропотов Ю.А. Анализ факторов снижения разборчивости речи в системах громкоговорящей связи // ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ. 2008. №1-3. С. 129-133.
11. Рудаков И.В., Романов А.С. Распознавание текстового изображения с учетом морфологии слова // Наука и образование: электронное научно-техническое издание. 2012. № 4. Режим доступа: http://technomag.bmstu.ru/doc/350020.html (дата обращения 10.1.2015).
12. Галкин В.А., Чернуха С.Н. Исследование быстродействия нейросетевого распознавателя подчерка // Наука и образование: электронное научно-техническое издание. 2011. № 12. Режим доступа: http://technomag.bmstu.ru/doc/280351.html (дата обращения 10.01.2015).
13. Гапочкин А.В. Нейронные сети в системах распознавания речи // Science Time. 2014. №1. С. 29-36.
14. Нейронные сети для любопытных программистов. Режим доступа: http://habrahabr.ru/sandbox/76908/ (дата обращения 10.01.2015).