ПРОБЛЕМЫ ТЕХНИЧЕСКОГО ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

ПРИМЕНЕНИЕ ЦИФРОВОЙ ОБРАБОТКИ ГОЛОСОВЫХ СИГНАЛОВ ДЛЯ УЛУЧШЕНИЯ РАСПОЗНАВАНИЯ РЕЧИ

А.А. Дмитриев, Д.А. Дмитриев
Алтайский государственный университет, г. Барнаул
email: dmitriev@asu.ru

Аннотация. В работе предложен метод предварительной обработки голосовых аудиозаписей, полученных по телефонной линии связи при разговоре пользователя с голосовым помощником. Для распознавания слов в работе использован программно-аппаратный комплекс, построенный на основе программного обеспечения Kaldi. Показано, что полученные голосовые сигналы могут быть искажены шумом, связанным с работой устройств телефонной сети. Поэтому для надежного распознавания слов в записанной речи применена предварительная фильтрация сигнала. Для выполнения обработки использован полосовой фильтр. Применение цифровой фильтрации позволило улучшить качество записи и уменьшить ошибку в распознавании отдельных слов в записанных сигналах.

Ключевые слова: распознавание речи, цифровая фильтрация, обработка сигналов.

Введение. Системы распознавания и синтеза речи находят сегодня широкое применение в бизнес-процессах различных коммерческих организаций. Обычно данные системы используются на телефонных линиях поддержки пользователей в виде голосовых помощников [1-2]. Голосовые помощники поддерживают речевой диалог с пользователем для выяснения причин обращения на горячую линию. Нередко на системы распознавания и синтеза речи возлагают функции идентификации в речи отдельных слов, связанных с различными угрозами безопасности для организации. Примерами подобных систем являются голосовые помощники, разработанные ведущими компаниями IT-индустрии, такие как Google, Yandex и др. Программные продукты перечисленных компаний имеют широкий функционал для работы голосового помощника, основанный на нейронных сетях, а также инструменты для предварительной обработки голосовых сигналов. С другой стороны, определёнными достоинствами обладают системы распознавания и синтеза речи, построенные на основе свободно распространяемых продуктов, такие как CMU Sphinx, Kaldi, RWTH ASR [1,3]. Данные системы также построены на математическом аппарате нейронных сетей. Точность распознавания речи при их использовании зависит от проведенного дообучения нейронной сети и предобработки записанного речевого сигнала. Для предобработки сигналов обычно применяются методы цифровой фильтрации. На этом этапе широко используется частотная фильтрация и более сложные методы, основанные на алгоритмах Фурье-анализа и вейвлет-фильтров. При этом выбор фильтра зависит от конкретного способа регистрации речевого сигнала.

В настоящей работе предложен подход к обработке записанного голосового сигнала с помощью цифровых фильтров для уменьшения ошибок при распознавании речи, выполненной предобученной нейронной сети программного обеспечения Kaldi.

Описание программно-аппаратного комплекса. Для регистрации и обработки речевых сигналов в работе использован программно-аппаратный комплекс, представленный на рис. 1.

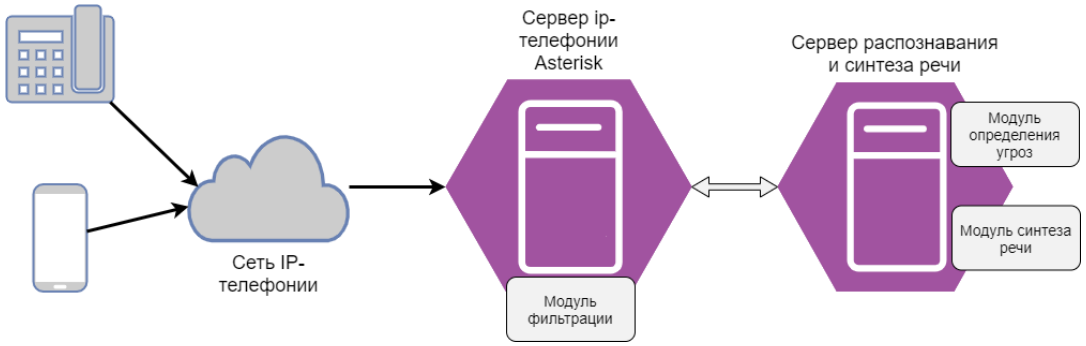

Рисунок.1. Схема модулей программно-аппаратного комплекса для регистрации звонков и распознавания и синтеза речи.

Как следует из рис. 1, программно-аппаратный комплекс состоял из двух вычислительных серверов. Сервер приема звонков был подключен к телефонным VOIP-сетям организации. На этот сервер поступали входящие звонки пользователей, обратившихся на горячую линию поддержки. Для регистрации телефонных звонков на данный сервер было установлено программное обеспечение для работы ip-телефонии Asterisk [4]. После регистрации телефонного звонка отдельным программным модулем проводилась его фильтрация для улучшения качества, а затем голосовое сообщение перенаправлялось на сервер распознавания речи. Для быстрой передачи голосовых данных между сервером Asterisk и сервером распознавания речи была организована отдельная высокоскоростная линия связи.

Сервер распознавания и синтеза речи использовался для обработки речевого потока и его преобразования в текстовые сообщения. На этом сервере для распознавания речи было установлено программное обеспечение Kaldi [5].

Программное обеспечение Kaldi состоит из набора программных инструментов для вычисления информативных параметров голосового сигнала и распознавания речи. В качестве информативных параметров, характеризующих наличие в сигнале отдельных звуков речи, при обработке сигнала рассчитывались нормализованные мел-кепстральные коэффициенты [5]. Для выделения шумовых составляющих в сигнале и особенностей речи говорящего применялся математический аппарат вычисления i-векторов [6]. Общий вектор признаков, состоящий из мел-кепстральных коэффициентов и i-векторов, обрабатывался предобученной нейронной сетью с временной задержкой [7]. На выходе нейронной сети формировались отдельные текстовые эквиваленты звуков букв и слогов, присутствующих в голосовом сигнале [5]. Для преобразования звуков в буквы с целью объединения их в слова и построения в виде текста законченных фраз, произнесенных пользователем, применялся математический аппарат теории автоматов [8].

Выбор Kaldi был основан на проведенном нами анализе публикаций, в которых указывалось, что Kaldi обеспечивает высокую точность распознавания речи на русском языке среди других свободно распространяемых программных продуктов с аналогичными функциональными возможностями [1,5]. В результате обработки речевого потока системой Kaldi создавался набор распознанных текстовых слов. Затем каждое слово анализировалось отдельным программным модулем для выявления слов и словосочетаний, содержащих различные угрозы и оскорбления в адрес организации. В случае обнаружения таких слов программный модуль автоматически информировал службы безопасности организации средствами электронной почты.

Одновременно с анализом слов модулем определения угроз распознанные текстовые слова передавались в программный модуль синтеза речи. Данный модуль обеспечивал классификацию распознанных слов для определения тематики

разговора и составления ответных речевых фраз. Таким образом, модуль синтеза речи использовался для поддержания непрерывного диалога с пользователем.

Использование нейронной сети из пакета программ Kaldi позволило эффективно распознавать отдельные слова речевом диалоге и преобразовывать их в текстовые записи. Однако некоторые слова в речевом потоке были распознаны с ошибками, которые состояли из искажений в виде пропущенных или добавленных букв. Проведенный анализ голосовых записей звонков пользователей, для которых система распознавания допускала ошибки, показал, что эти записи были искажены фоновым шумом, связанным с работой телефонных аппаратов. В связи с этим, в настоящей работе было предложено использовать цифровую фильтрацию для предварительного улучшения качества голосовых данных.

Цифровая обработка речевых сигналов. Для обработки речевые сигналы были предварительно записаны с частотой дискретизации 8 кГц и сохранены в формате wav. Выбор такой частоты дискретизации обусловлен тем, что современные телефонные сети общего пользования спроектированы для передачи звуков в диапазоне частот 300-3500 Гц [4]. Для сигналов, в которых при распознавании речи были допущены ошибки, были построены спектрограммы в координатах «амплитуда коэффициентов A – частота F, Гц», как показано на рис 2a.

Как следует из рис. 2a, в спектре обрабатываемого сигнала присутствовали шумовые составляющие на частотах, выходящих за рабочий диапазон в 300-3500 Гц. Происхождение данных шумов, как было отмечено выше, было связано с работой телефонных устройств. Для подавления искажающего шума был применен полосовой ких-фильтр [9]. Необходимая крутизна амплитудно-частотной характеристики фильтра на частотах среза обеспечивалась высоким порядком фильтра, в котором использовались 120 коэффициентов фильтрации. Коэффициенты рассчитывались на основе окна Хемминга. На рис 2b в виде частотного спектра представлен результат обработки сигнала.

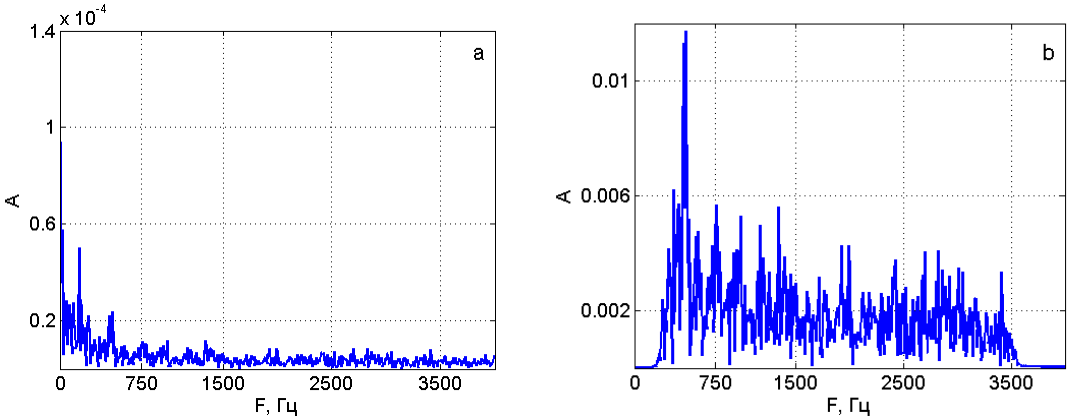

Рисунок 2. Спектр записанного речевого сигнала:
a – спектр сигнала до обработки, b – спектр сигнала после проведенной фильтрации.

Рис. 2b показывает, что в результате фильтрации, были удалены шумовые составляющие, лежащие вне рабочей части спектра. Отметим, что применение предложенной обработки не приводило к значительному уменьшению громкости звуков в аудиозаписи и искажению произносимых слов. Обработанные фильтром речевые сигналы, а также записанные сигналы без обработки затем использовались для распознавания речи.

Для количественной оценки качества распознавания речи в настоящей работе использован параметр WER [1, 10]. Параметр WER является метрикой, характеризующей количество ошибок в словах при распознавании речи, и вычисляется согласно выражению 1.

$$WER = \frac{S+D+I}{N} \quad (1)$$

Здесь S обозначает число слов с заменой, удалением или вставкой букв, коэффициент D характеризует количество удаленных слов из фразы, I – количество новых слов, неправильно добавленных в речевую фразу, а N – общее число слов. Результаты расчета WER для записанных нами речевых сигналов в исходном виде и после проведения фильтрации показаны в таблице 1.

Таблица 1. Рассчитанные значения WER при распознавании речи нефильтрованных и фильтрованных сигналов, в %.

$WER_{исх}$	$WER_{фил}$
21.1%	20.8%

Как следует из таблицы 2, распознавание фильтрованных речевых сигналов происходило с меньшей ошибкой $WER_{фил}$ = 20.8%, чем при распознании сигналов не прошедших фильтрацию $WER_{исх}$ = 21.1%. Полученный результат указывал на то, что проведенная фильтрация сигнала способствовала более точному описанию особенностей амплитудно-частотных характеристик сигнала с помощью информативных параметров, вычисленных программным обеспечением Kaldi. Применение описанной выше предварительной обработки записанных голосовых сигналов приводило к более качественному распознаванию отдельных слов. Однако, для существенного уменьшения ошибки WER при распознавании слов представленный метод фильтрации, по-видимому, необходимо сочетать с дообучением используемой нейронной модели.

Заключение. Надежное распознавание речи программным обеспечением, реализованным на основе нейронных сетей, зависит как от сложности применяемой нейронной сети, используемых информативных параметров, так и от способов предобработки голосовых сигналов. В работе описан подход к предварительной обработке сигналов цифровым полосовым фильтром. Проведенная обработка способствовала устранению шумовых составляющих сигнала и улучшала точность распознавания отдельных слов в речи.

Полученные результаты показали, что предложенный метод фильтрации может использоваться, как один из способов для предварительного улучшения сигналов в программно-аппаратных комплексах для распознавания речи.

Библиографический список.

1. Беленко М.В., Балакшин П.В. Сравнительный анализ систем распознавания речи с открытым кодом // Международный научно-исследовательский журнал. – 2017. - №4(58). – С. 13-18.
2. Jha M. Improved unsupervised speech recognition system using MLLR speaker adaptation and confidence measurement // V Jornadas en Tecnologıas del Habla (VJTH'2008). – 2008. – P. 255-258.
3. Ravanelli M., Parcollet T., Bengio Y. The pytorch-kaldi speech recognition toolkit // 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2019. – P. 6465-6469.
4. Брайант Р., Медсен Л., Меггелен Д. В. Asterisk: окончательное руководство // O'Reilly Media, 2013. – 641 p.
5. Povey D., Ghoshal A., Boulianne G. The Kaldi Speech Recognition Toolkit // IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. – 2011. – P. 1-4 .
6. Берзинь А.У. Применение i-векторов для автоматизированного определения уровня близости языков // Труды ИСП РАН. – 2019. – Т. 31. Вып. 5. – С. 153 - 164.
7. Peddinti V., Povey D., Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts // Interspeech. – 2015. – P. 3214-3218.

8. Georgescu A.-L., Cucu H., Burileanu C. Kaldi-based DNN Architectures for Speech Recognition in Romanian // 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). – 2019. – P. 1-6.
9. Сергиенко А.Б. Цифровая обработка сигналов // СПб. : Питер, 2002. – 608 с.
10. Карпов А.А., Кипяткова И.С. Методология оценивания работы систем автоматического распознавания речи // ИЗВ. ВУЗОВ. ПРИБОРОСТРОЕНИЕ. – 2012. – Т. 55, № 11. – С. 38-43.

USING OF DIGITAL VOICE SIGNAL PROCESSING TO IMPROVE SPEECH RECOGNITION

A.A. Dmitriev, D.A. Dmitriev
Altai state university, Barnaul
email: dmitriev@asu.ru

Annotation. The paper proposes a method for pre-processing voice audio recordings received over a telephone line during a conversation between a user and a voice assistant. For speech recognition, a hardware-software complex built on the basis of Kaldi software was used in the work. It is shown that the received voice signals can be distorted by noise associated with the operation of telephone network devices. Therefore, for reliable recognition of words in the recorded speech, a preliminary filtering of the signal was applied. A band pass filter was used to perform the signal processing. The use of digital filtering made it possible to improve the quality of the recordings and reduce the error in recognizing individual words in the recorded signals.

Keywords: speech recognition, digital filtering, signal processing.