DOI 10.37882/2223–2966.2022.05.07

ОБРАТНАЯ ЗАДАЧА РАСПОЗНАВАНИЯ РЕЧИ

Гутенков Роман Леонидович
Аспирант, Российский Технологический Университет МИРЭА
rlggut@mail.ru

Аннотация. Данная работа посвящена обратной задаче по нахождению участка речи в сигнале. Исходный сигнал гарантированно содержит речь, однако неизвестно, каким кодеком он был закодирован. Основная цель — сформировать метод определения кодека и параметров предварительной обработки, таких как инверсия, разворот байта и кадра, которым изначально был кодирован сигнал. Рассмотрено несколько часто используемых детекторов речевой активности, а именно метод расчета энергии на участке сигнала и спектральный метод. На их основании сформировано предположение о том, что для определения кодека нужно использовать один из детекторов речевой активности, так как сигнал сразу начинается с речи. Используя один из детекторов и введя численную оценку, можно определить, какой кодек и какие параметры предварительной обработки требуется использовать, чтобы корректно декодировать сигнал. После чего потребуется провести проверку метода на разных сигналах. Как итог, в работе сформулирован общий подход для определения нужного кодека при условии, что сигнал гарантированно содержит речь.

Ключевые слова: распознавание речи, методы распознавания речи, преобразование Фурье, спектрограмм, детектор речевой активности.

INVERSE PROBLEM OF SPEECH RECOGNITION

R. Gutenkov

Summary. This work is devoted to the inverse problem of finding a segment of speech in a signal. The original signal is guaranteed to contain speech, but it is not known which codec it was encoded with. The main goal is to form a method for determining the codec and pre-processing parameters, such as inversion, byte and frame reversal, with which the signal was originally encoded. Several commonly used speech activity detectors are considered, namely the method for calculating the energy in the signal section and the spectral method. Based on them, an assumption was made that one of the speech activity detectors should be used to determine the codec, since the signal immediately begins with speech. Using one of the detectors and entering a numerical estimate, you can determine which codec and which pre-processing parameters need to be used in order to correctly decode the signal. After that, it will be necessary to test the method on different signals. As a result, the paper formulated a general approach for determining the required codec, provided that the signal is guaranteed to contain speech.

Keywords: speech recognition, speech recognition methods, Fourier transform, spectrograms, speech activity detector.

Введение

Определение участков речевой активности в звуковом сигнале является достаточно актуальной проблемой в связи с ростом спроса на голосовое управление, не упоминая тонкости обычного шумоподавления в телефонном разговоре. Проблемы возникают как из-за речевых специфик, так и из-за различных шумов, которые наложились на речевой сигнал. На данный момент уже существуют множество систем распознавания речи, основанные на анализе энергии участка сигнала, построение спектрограммы через преобразование Фурье или других аналитических моделях.

Однако все это применимо только в том случае, когда исходный сигнал правильно декодирован. Если рассматривать исследуемый сигнал, который неправильно декодировали, то результат будет достаточно непредсказуемым. К примеру, участок сигнала, который при правильно кодировке на самом деле был паузой, станет похож на шум, а речевой участок, наоборот, станет ближе к паузе. Помимо очевидно не корректной работы, детекторы речевой активности (VAD), которые опираются на паузы в речи, перестанут работать корректно.

Проблема не точного декодирования является довольно острой. Рассмотрим пример сигнала, который закодирован при помощи кодека g711a, или, иначе aLaw. Большинство сигналов, закодированных с его помощью, можно декодировать и с помощью g711u (он же uLaw) так, чтобы было слышно в сигнале речь (рис. 1). Но при этом образуется шум, с которым придется дополнительно справляться детекторам речевой активности. Чтобы избежать данной проблемы, требуется изначально быть уверенным, что сигнал, в котором ищут речь, правильно декодировали.

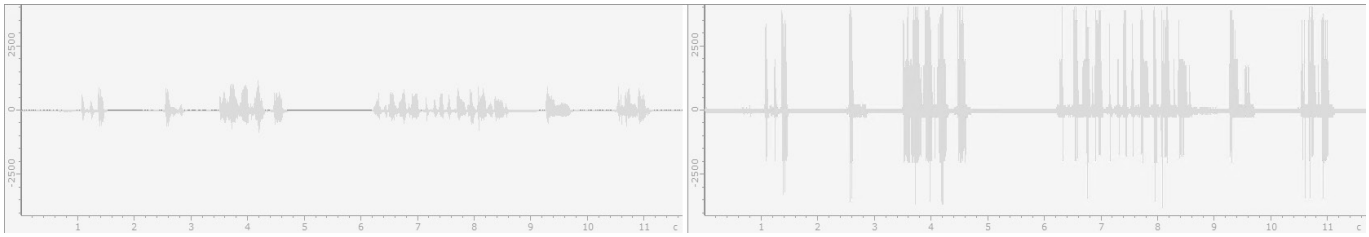

Рис. 1. Закодированный сигнал aLaw, декодированный кодеками aLaw (слева) и uLaw (справа)

Попробуем посмотреть на проблему с другой стороны. Пусть заранее известно, что данный сигнал содержит речь в его начале. Что тогда требуется? Какими критериями нужно руководствоваться, чтобы правильно определить кодек, которым закодировали сигнал?

Данная задача практически является обратной к задаче по нахождению речевого участка в сигнале. Заранее известно наличие речи в сигнале, но не известен кодек, с помощью которого данный сигнал был закодирован.

В статье предлагается подход к определению кодека и параметров предварительной обработки, которыми закодирован сигнал при условии, что начало сигнала является участком речи.

Литературный обзор

Данная задача напрямую не встречается литературе последних лет, что может свидетельствовать о новизне поставленной задачи. На практике сигнал уже подготовлен и не требуется проверять, правильно ли он был декодирован.

Обратная задача, где известно, что сигнал начинается с речевого участка, но декодирован некорректно, встречается достаточно редко. Такая задача встречается только при условии, что пользователь забыл или потерял некоторые параметры декодирования исходного сигнал. Вполне вероятно, что решение обратной задачи на данный момент реализуется перебором напрямую, без математического обоснования.

Однако про прямую задачу, нахождения участков речи в сигнале, написано много статей. Так как в дальнейшем будут использованы методы нахождения участков речи в сигнале, стоит подробнее разобрать ситуацию с детекторами речевой активности, сложившуюся на данный момент.

Крайне простым по вычислительным критериям, является детектор, основанный на энергии участка сигнала. Подсчитав сумму квадратов колебаний, поделив на длину временного участка и сравнив с некоторым эталоном можно сделать некоторые выводы. Если подсчитанная энергия превышает эталон, то считать участок, на котором была подсчитана энергия, участком с речью. Иначе участок является паузой в речевом сигнале.

Преимуществами данного способа являются простота реализации и быстрота выполнения. Однако недостатков данного подхода гораздо больше: чувствительность к количеству пауз в речи на единицу времени, тихая речь может смешаться с шумом, при неправильном декодировании сигнал может весь стать громким шумом.

Дальнейшие усовершенствования данного метода ведут к анализу изменения энергии во времени, сравнении скачков энергии с предполагаемой моделью изменений и так далее. Но и это не решает полностью проблему ложного срабатывания. Ведь если участок сигнала будет звучать как журчание или нечто схожее, то даже измененный подход не сможет качественно разобраться с данной ситуацией.

Наиболее подходящим применением подсчета энергии в области распознавания речи является нахождение участков шума или потенциальной речи во всем сигнале для дальнейшего анализа найденного участка.

Другой метод детектирования голосовой активности — число нулевых переходов (ZCR). Эта величина показывает грубую оценку спектральных свойств акустического сигнала. Чем меньше значение ZCR, тем выше шанс того, что данный участков содержит речь, чем шум, для которого данное значение является случайным.

Недостатками данного метода является недостаточно эффективное распознавание речи, поскольку музыка и некоторые шумы могут определяться как речь. Или, наоборот, не вокальная речь определяется как шум.

Как развитие спектрального подхода, стоит упомянуть быстрое преобразование Фурье и построение спектрограммы. Разложив сигнал и построив амплитудную

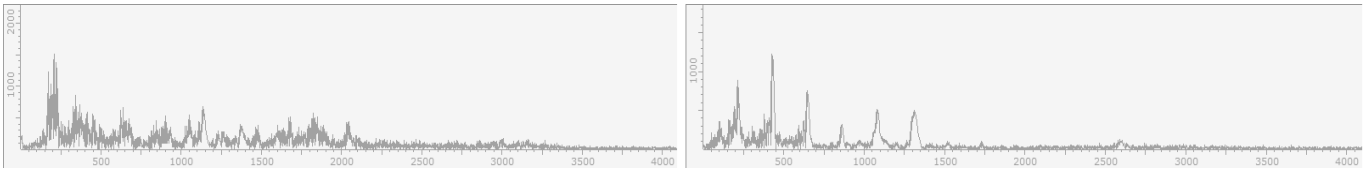

Рис. 2. Спектрограммы речевых участков

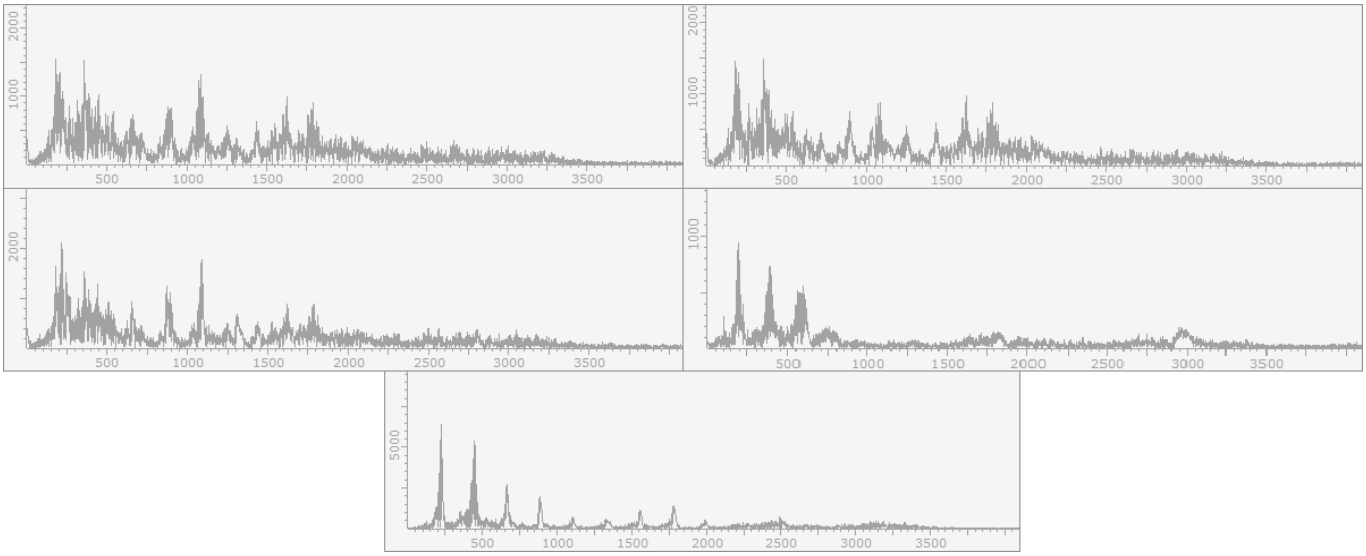

Рис. 3. Спектрограммы речевых участков, декодированных сверху вниз и слева направо: aLaw, uLaw, g726_24, g726_32, g729

спектрограмму, можно убедиться, что в речи чаще преобладают высокие частоты. Или использовать оконное преобразование Фурье. Тогда основным ориентиром данного детектора будет плотность функции распределения гармоник. Из недостатков данного метода можно отметить большее время работы, сравнивая с предыдущими методами. Проблему выбросов, определения временного отрезка для расчётов и создание эталонной спектрограммы тоже стоит учесть.

Существуют другие речевые признаки и детекторы, построенные на них, однако в целом ни один из них не является универсальным. Каждый подход лучше или хуже в зависимости от исходных данных, поэтому в общем случае для обнаружения речевых участков применяют комбинацию вышеперечисленных методов или их модифицированные вариации.

В работе выбор будет остановлен на применении быстрого преобразования Фурье, что объясняется большей надежностью, сравнивая с другими методами. Комбинированный метод не рассматривался, чтобы не перегружать первоначальное приближение в решении поставленной проблемы.

Прежде, чем перейти к рассмотрению непосредственно задачи, стоит также уточнить некоторые моменты, касающиеся кодеков. Например, стоит учитывать то, что кодеки aLaw и uLaw схожи между собой, о чем уже упоминалось ранее.

Также существует такой термин, как речевой кодек — кодек, который используется в основном только для речи, такой как g729. Данный кодек отличается от тех же g711a и g711b тем, что блок его кодирования составляет 80 бит по сравнению с 8 битами для g711. Дополнительно отличие кроется в подходе, ведь кодек g729 Annex B содержит индикаторы тишины, что позволяет дополнительно сжать участки тишины в речи. Поэтому использовать методы, которые опираются на особенности кодеков не стоит, чтобы не умалять общности подхода.

Материалы и методы

Прежде, чем перейти к методам, нужно четко сформулировать цель исследования. Цель — сформулировать метод, с помощью которого можно определить, каким кодеком и с какими параметрами предварительной

Таблица 1. Нормированное значение 8ми участков спектрограмм из рисунка 3 и их среднее значение.

aLaw	100	60	59	58	31	31	25	19
uLaw	100	59	58	57	30	30	25	19
G726_24	100	63	60	44	22	22	28	22
G726_32	100	61	62	58	31	31	29	22
G729	100	60	23	37	26	26	24	26
Среднее	100	60,6	52,4	50,8	28	28	26,2	21,6

Таблица 2. Результаты применения метода к расчетным данным.

Сигнал	Оценка	Инверсия	Разворот байта	Разворот кадра	Кодек
aLaw	44	Нет	Нет	Нет	aLaw
	44	Нет	Да	Да	aLaw
uLaw	42	Нет	Нет	Нет	uLaw
	59	Нет	Нет	Нет	aLaw
G726_24	35	Нет	Да	Да	G726_24
	36	Нет	Нет	Нет	G726_24
G726_32	43	Нет	Нет	Нет	G726_32
	43	Нет	Да	Да	G726_32
G729	66	Нет	Нет	Нет	G729
	131	Нет	Нет	Да	G729

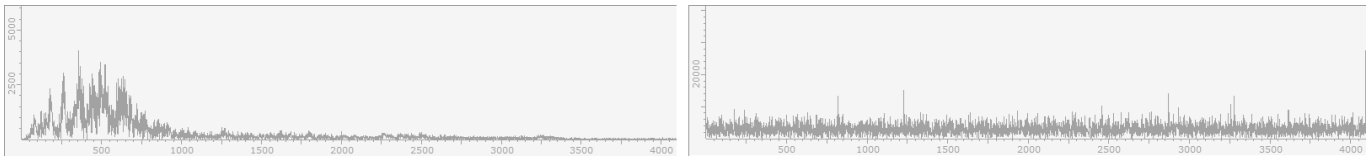

Рис. 4. Изначальный сигнал g726 32кб/с декодирован кодеком g726 (слева) и сигнал с кодеком g729 декодирован кодеком g726_32 (справа)

обработки (инверсия, разворот кадра, разворот байта) нужно декодировать исходный сигнал при условии, что он начинается с речевого участка, длиной не менее 1.2 с. Последнее уточнение требуется для использования преобразования Фурье, о чем написано ниже.

Далее будут рассматриваться такие кодеки, как aLaw, uLaw, g729, g726 24 и 32 кб\с. Рассматриваемые ниже методы не опираются на уникальность какого-либо из кодеков, что позволяет в полной мере описывать общий подход для любого набора кодеков, используя выбранную группу.

Если известно, что исходный сигнал в его начале является участком речи, то допустимо перебрать кодеки из установленного списка, декодировать ими исследуемый сигнал, после чего провести анализ на наличие речи в декодированном сигнале.

Данная цепочка рассуждений приводит к использованию стандартных детекторов речевой активности для определения корректного кодека для декодирования.

Из вышеперечисленных в литературном обзоре детекторов речевой активности был выбран метод построения спектрограмм с применением преобразования Фурье. Подробнее про выбор и сам метод описано выше.

Для определенности будем проводить преобразование Фурье на отрезке примерно в 1 секунду. При усло-

вии, что все декодируется в стандартный формат PCM, для этого потребуется брать 8000 отсчетов из сигнала. Если применять быстрое преобразование Фурье, то лучше использовать 8129 отсчета, так как это степень двойки.

Построив спектрограммы нескольких речевых участков, сформулируем некоторую идеальную модель распределения спектрограммы, характерную только для речи. В дальнейшем сравнивая модель с результатом декодирования, можно будет сделать вывод о том, правильно ли был декодирован сигнал.

Для создания модели потребуется использовать некоторое количество, не менее 4х, спектрограмм различных речевых участков. Во всех случаях требуется использовать одинаковую продолжительность. Чтобы не сравнивать целиком спектрограммы, целесообразнее их разбивать на участки, после чего высчитывать среднее значение на участках и сравнивать уже их. Отталкиваясь от полученных спектрограмм, можно сделать вывод, что разбиение на 8 равных участков будет достаточным. Каждый из участков потребуется нормировать, так как это необходимо, чтобы нивелировать проблему различной громкости речевых участках.

Подсчитав для каждой спектрограммы из рисунка 3 суммы участком и нормируя их, получаем следующие значения

Подсчитывая отклонения между одинаковыми участками спектрограммы предполагаемого речевого участка и эталона, можно делать выводы о корректности декодирования, вплоть до численного ранжирования наиболее подходящих кодеков.

Результаты

Как итог, общими словами алгоритм действий для определения корректного кодека для декодирования сигнала будет выглядеть следующим образом:

Исходный сигнал по очереди декодируют некоторым заранее определенным набором кодеков. Далее, в полученном сигнале с помощью подсчета энергии на участке определяется участок с наибольшей энергией, чтобы провести на нем преобразование Фурье и построить спектрограмму. После чего, вычислив отклонение по модулю от заранее составленного эталона (метод описан выше), ранжировать полученные результаты всех кодеков. Кодек с наименьшим числом должен быть наиболее корректно декодирующим исходный сигнал.

Чтобы проверить, правильный ли был выбранный метод, нужно проверить модель на тех же сигналах, на которых была сформирована модель. Потребуется декодировать сигналы с различным сочетанием параметров предварительной обработки и всеми кодеками из выбранной группы. Если для каждого из сигналов лучшим результатом будет тот же кодек, что и подразумевался, то можно считать, описанный выше алгоритм достаточно точно определяет кодек, которым был закодирован изначально сигнал.

Была составлена программа, которая перебирает все варианты для каждого сигнала, подсчитывает оценку и составляет таблицу с результатами. Чтобы не приводить полностью все таблицы с вариантами (для каждого сигнала проверяется каждый кодек из списка с перебором всех 3х параметров предварительной обработки, что суммарно превращается в 5*3*3*3=45 вариантов для каждого сигнала), приведем только первые две строки.

Обсуждение

Как уже отмечалось ранее, если использовать кодек aLaw вместо uLaw речь будет слышна, но она будет с шумом. Результат того, что применение одного вместо другого оказалось так близко к правильному кодированию первого сигнала кроется в нормировании участков, ведь и шум тоже нормируется при этом.

Ситуации, когда варианты с одновременным использованием разворота байта и разворота кодека близки к вариантам без них, тоже достаточно понятны. Такое бывает у кодеков, в которых ровно 8 бит на кадр, либо получается так из-за других особенностей, как это вышло с G726_32. По итогу, хоть и с шумами, но вариант с одновременным разворотом и байта, и кадра совпадал с вариантом без них.

Заключение

Данный подход работает на контрольных примерах и претендует на потенциальное решение поставленной цели работы, но требует дальнейшей проверки. Описанный выше метод не лишен проблемных мест, таких как правильно подобранный эталон и вырожденные случаи. Однако с ними можно разобраться при дальнейшей апробации метода такими способами, как увеличение выборки для эталона, введя систему весов для областей или даже введения двух эталонов.

Благодарности

Описанные выше методы и идеи были навеяны коллегами по работе и по институту, за что им крайне признателен.

ЛИТЕРАТУРА

1. Волченков В.А., Витязев В.В. Методы и алгоритмы детектирования активности речи // Цифровая обработка сигналов. 2013. № 1. С. 54–60.
2. Гаврилович Н.В. Методы распознавания речи и их классификация // Таврический научный обозреватель. 2016. № 6. С. 206–212
3. Кравцов С.А. Исследование работы детектора речевой активности в задаче идентификации диктора // Радиотехнические и телекоммуникационные системы. 2015. № 4 (20). С. 61–68.
4. Кухтинова М.С., Позолотина Н.А., Трубин В.Г. Системы распознавания речи // Автоматика и программная инженерия. 2014, № 2(8), С. 46–47.
5. Панова А.А., Яковенко А.А. Методы детектирования голосовой активности // Системный анализ в проектировании и управлении. 2019. С. 397–403
6. Ramírez J., Górriz J.M., Segura J.C. Voice activity detection. Fundamentals and speech recognition system robustness // Robust Speech Recognition and Understanding. Vienna: I-TECH Education and Publishing, 2007. P. 1–22.

Журнал «Современная наука: актуальные проблемы теории и практики»

МИРЭА — Российский технологический университет