Научная статья
Original article
УДК 004.934

МЕТОДЫ ДЕТЕКТИРОВАНИЯ ЗВУКОВОЙ АКТИВНОСТИ

METHODS FOR SOUND ACTIVITY DETECTION

Гутенков Роман Леонидович, аспирант, Российский Технологический Университет МИРЭА, г. Москва

Gutenkov Roman Leonidovich, post-graduate student, Russian Technological University MIREA, Moscow

Аннотация

Работа посвящена рассмотрению уже известных методов детектирования речевой активности и некоторым особенностям их применения. Исследуется проблема первичной подготовки сигнала к поиску речевых участков в нем. Использование метода подсчета энергии позволяет отбирать участки, которые гарантированно содержат звуковую активность, шум и речь, и отбрасывать те, в которых речь точно отсутствует. Если предварительно произвести подобный отбор, то суммарная расчетная сложность задачи уменьшится. Также рассмотрен метод, аналогичный подсчету мощностей, но требующий сравнительно меньше вычислительных действий, который справляется с описанной выше задачей. Произведено сравнение рассматриваемых методов и их оценка эффективности.

Annotation

The work is devoted to the consideration of already known methods for detecting speech activity and some features of their application. The problem of the initial signal preparation for the search for speech segments in it is investigated. The use of the

energy counting method makes it possible to select areas that are guaranteed to contain sound activity, noise and speech, and discard those in which speech is definitely absent. If such a selection is made in advance, then the total computational complexity of the problem will decrease. Also considered is a method similar to power calculation, but requiring relatively less computational steps, which copes with the task described above. Comparison of the considered methods and their efficiency estimation is made.

Ключевые слова: обработка сигналов, детектор голосовой активности, краткосрочная энергия, распознавание речи, машинных слух

Keywords: signal processing, voice activity detector, short-term energy, speech recognition, machine hearing

В настоящее время продолжают активно развиваться технологии анализа речевых сигналов. Данный вопрос является ключевым для классификации акустических событий, фильтрации сигнала, голосовых интерфейсов. В самом сигнале помимо речи, могут присутствовать посторонние шумы, гудки и прочие акустические события, но прежде, чем обрабатывать и анализировать сигнал, необходимо выявить участки, на которых данные события могут присутствовать. Потому можно считать, что любой анализ речевого сигнала начинается с определения участков с потенциальным присутствием речи, чтобы в дальнейшем использовать детекторы голосовой активности (Voice Activity Detection, VAD).

Данное предложение оценить просто, так как после выделения отдельных участков сигнала с акустическими событиями, уменьшается вычислительная сложность анализа всего сигнала по сравнению с анализом исходного варианта. Что в свою очередь увеличивает скорость обработки сигнала.

Опираясь на многочисленные исследования, можно утверждать, что в речи может содержаться до 50% пауз, а в диалоге данный показатель может доходить до 70% от всего сигнала. Из-за чего проблема первичного разбора сигнала стоит достаточно остро.

На сегодняшний день известно множество методов детектирования речи, однако разработка новых методов продолжается. В автоматизированных системах

обработки речи детектор голосовой активности позволяет разделить сигнал на речевые и неречевые участки. Задача заключается в подборе подходящего критерия и самого детектора, которые позволят с высокой долей вероятности определить, что данный участок сигнала содержит речь. В случае неверно принятого критерия может произойти потеря части речевых данных.

Считается, что речевые участки акустического сигнала в сочетании с линейными неречевыми источниками, обладают некоторыми постоянными признаками, по которым можно их выделить. Существуют разнообразные методы параметризации речевых сигналов и самый распространенный из них – мощность сигнала.

Понятие мощности для сигнала на участке $[t1; t2]$ определяется по формуле

$$E = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} s^2(t)\, dt$$

, где *s(t)* обозначает значение сигнала в момент времени *t*. При превышении некоего порога, можно считать, что на данном интервале присутствует речь или другая звуковая активность, отличная от фонового шума.

Из преимуществ данного метода нужно отметить простоту реализации и относительную быстроту выполнения по сравнению со спектральными методами. Однако и не обойдется без недостатков данного метода. При неверно выбранной длине исследуемого участка и большого количества пауз в речи на данном участке детектор может сработать некорректно. Также присутствует проблема с тихой речью, что может смешаться с шумом. Как способ решения, можно указать динамически изменяемый порог, который не просто задан изначально, но и который будет изменяться, в зависимости от сигнала. Но данный подход усложнить данный метод и уберет тем самым одно из очевидных достоинств – простоту реализации и скорость выполнения.

Следующим по частоте упоминания в научных работах можно считать метод нулевых переходов (Zero Crossing Rate или ZCR). Этот параметр делает грубую оценку спектральных свойств. Данный метод построен на предположении, что вокализированные участки обладают гораздо меньшим значением ZCR, чем

участки с шумом. Применение данного метода имеет некоторые ограничения – некоторые шумы и музыка также будут определяться как речь, из-за малого значения ZCR. Одновременно с этим, не вокализированные участки речи будут считаться шумом, так как исходное предположение метода опирается именно на вокализированную речь.

Более фундаментальным детектором речевой активности являются спектральные детекторы. Например, применив на участке быстрое преобразование Фурье можно построить амплитудную спектрограмму. На основе полученной спектрограммы можно сделать предположение о характерной только для речи плотности функции распределения гармоник. Или провести оконное преобразование Фурье и опираться уже на спектр оконной функции.

Для данных подходов характерна проблема определения размера исследуемой длины сигнала и большая, в сравнении с вышеперечисленными методами, вычислительная сложность. Однако и результаты будут качественнее: в большем количестве случаев корректно отделят речевые от не речевых участков.

Существуют и другие детекторы речевой активности, основанные на других речевых признаках, но у них нет полной универсальности. В зависимости от специфики исходных сигналов, часть из них будет показывать себя лучше, чем другие. Но общее у них всех одно – если применять их не ко всему сигналу, а только к участкам, где гарантированно не может быть пауз, то это уменьшит общее время работы.

Чтобы отделить участки, в которых может быть речь, от условных пауз и шумового фона, воспользуемся методом подсчета энергии. Если подходить к данному методу с точки зрения отбора участком с речью и шумом от прочих, данный метод подходит больше остальных.

Для речевого сигнала стоит обозначить такое понятие, как монотонные области. По изменению энергии относительно предыдущих значений, можно четко разделить такие монотонные области речевого сигнала, как тишину, звуковую активность или однотонные звуковые сигналы (гудки, свисты и прочие).

Тишина от звуковой активности и монотонных сигналов отличается уровнем энергии. Звуковая активность бывает различной, как уже упоминалось ранее. Это может быть как речь, так и просто громкий шум. Однако наша цель отделить участки, которые заведомо не стоит рассматривать более сложными методами, от тех, в которых может находиться речь.

Если на исследуемом участке энергия сигнала поддерживается на почти одном и том же уровне, то можно утверждать, что данный участок относится к звуковому сигналу. При превышении порога можно утверждать о том, что в этих участках гарантированно содержится речь.

Реализация применения метода подсчета мощности совместно с выделением областей монотонности в общих чертах будет выглядеть следующим образом: исходный сигнал необходимо разбить на небольшие интервалы и посчитать на них энергию. Оценив уровень энергии на участках, можно сделать вывод об их принадлежности к одному из двух типов (подлежащих к дальнейшему рассмотрению или нет).

Выбор интервала является важной проблемой, так как с одной стороны стоит учесть, что от длины интервала будет зависеть, как много речевых данных будет обрезано. Для дальнейших рассуждений будем использовать интервал в 50 миллисекунды. Этого достаточно, чтобы уследить за переходом от паузы до речи и не достаточно, чтобы в речи было много пауз.

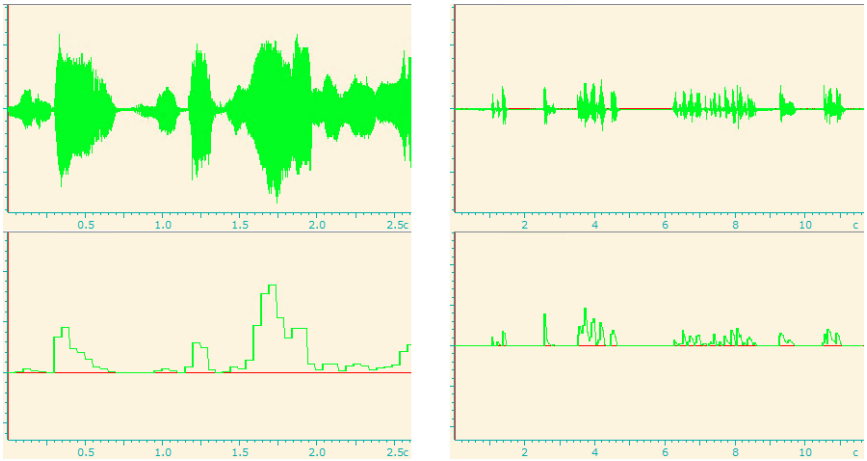

Рис.1. Сигналы и энергия сигналов, подсчитанная на интервале в 50 миллисекунд

Рассмотрим некоторый аналог мощности энергии сигнала для детектирования голосовой активности, а именно сумму модулей значений сигнала на участке $[t1; t2]$:

$$\frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} |s(t)|$$

Данная замена, по сути, подсчитывает площадь под графиком звукового сигнала.

Если площадь сохраняется между участками, то можно сделать вывод о том, что это монотонное звуковое событие. Если площадь под графиком меньше заданного порога, то можно утверждать о наличии только фонового шума, иначе этот участок может быть как шумом, так и речью.

Таким образом, можно сделать вывод, что для нахождения монотонных областей звукового сигнала достаточно исследовать площадь под звуковой разверткой данного сигнала.

Проведем сравнение графиков мощностей и суммы модулей на речевом сигнале. Для этого будем брать одинаковые небольшие интервалы (в 50 миллисекунд), считать на них мощность и сумму модулей. После чего сравним полученные результаты.

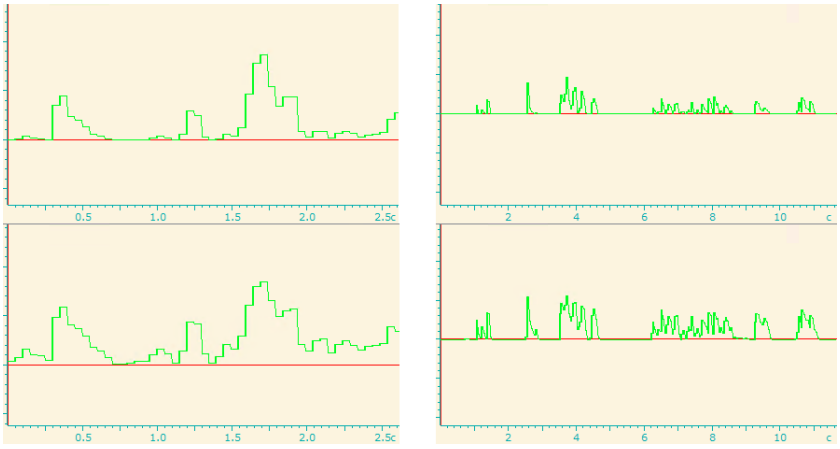

Рис.2. Сравнение использования суммы квадратов (верхний ряд) и суммы модулей (нижний ряд) значений сигналов в формуле подсчета энергии

Как можно заметить, для анализа областей монотонности достаточно использовать сумму модулей, вместо подсчета энергии на интервале.

Рассмотренный метод, основанный на подсчете суммы модулей значений сигнала, требует меньше вычислительных действий по сравнению с методом вычисления мощности сигнала. При этом данный метод не уступает по эффективности и целесообразности методу подсчета мощности для поставленной задачи – поиска монотонных областей. Из чего можно сделать вывод, что для определения участков, в которых может содержаться речь, достаточно использовать сумму модулей значений сигнала, вместо суммы квадратов.

Литература

1. Волченков В.А., Витязев В.В. Методы и алгоритмы детектирования активности речи // Цифровая обработка сигналов. 2013. №1. С. 54-60.
2. Кравцов С.А. Исследование работы детектора речевой активности в задаче идентификации диктора // Радиотехнические и телекоммуникационные системы. 2015. №4 (20). С. 61-68.
3. Кусков И. Э., Зотин А. Г. Сравнение алгоритмов обнаружения речевой активности // Актуальные проблемы авиации и космонавтики: сборник трудов V Международной научно-практической конференции, посвященной Дню космонавтики. Красноярск, 2019, том 2, С. 137-139.
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.

Literature

1. Volchenkov V.A., Vityazev V.V. Methods and algorithms for detecting speech activity // Digital signal processing. 2013. No. 1. pp. 54-60.
2. Kravtsov S.A. Investigation of the operation of the speech activity detector in the problem of speaker identification. 2015. No. 4 (20). pp. 61-68.

3. Kuskov I. E., Zotin A. G. Comparison of speech activity detection algorithms // Actual problems of aviation and astronautics: Proceedings of the V International Scientific and Practical Conference dedicated to the Day of Cosmonautics. Krasnoyarsk, 2019, volume 2, pp. 137-139.
4. Kukhtinova M.S., Pozolotina N.A., Trubin V.G., Speech recognition systems // Automation and software engineering. 2014, No. 2(8), pp. 46-47.
5. Panova A.A., Yakovenko A.A. Voice activity detection methods // System analysis in design and management. 2019. P.397-403
6. Ramírez J., Gorriz 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.

© Гутенков Р.Л., 2022. Научно-образовательный журнал для студентов и преподавателей «StudNet» №4/2022

Для цитирования: Гутенков Р.Л. МЕТОДЫ ДЕТЕКТИРОВАНИЯ ЗВУКОВОЙ АКТИВНОСТИ// Научно-образовательный журнал для студентов и преподавателей «StudNet» №4/2022