УДК 621.395
doi: 10.21685/2587-7704-2024-9-2-9

Open Access RESEARCH ARTICLE

Краткий обзор способов обнаружения речевой активности

Никита Николаевич Нефедов
Пензенский государственный университет, Россия, г. Пенза, ул. Красная, 40
nikita.nefyodow@mail.ru

Алан Казанферович Алимурадов
Пензенский государственный университет, Россия, г. Пенза, ул. Красная, 40
alansapfir@yandex.ru

Аннотация. Представлен краткий обзор способов обнаружения речевой активности, которые успешно используются в зашумленных условиях с различным соотношением сигнал/шум. Обнаружение речевой активности является процессом идентификации сегментов речи в непрерывном аудиопотоке и используется для сокращения вычислительных операций. Рассматриваются способы, основанные на гармоничности и частоте модуляции, которые обеспечивают хорошие результаты в условиях сильно зашумленной обстановки. Представлены результаты экспериментов способов обнаружения речевой активности на основе оценки гармоничности и частоты модуляции в рамках задачи обнаружения речи для управления автономным погрузчиком. В результате было показано, что комбинирование гармоничности и частоты модуляции позволяет повысить достоверность обнаружения речевой активности. Проведен обзор системы обнаружения речевой активности на основе SUB-BAND, которая использует максимальные значения отношения сигнал/шум поддиапазонов в качестве детекторных функций. Экспериментальные результаты показывают, что предложенный метод достигает лучших показателей по сравнению с обычными ETSI AMR VAD в базе данных NOISEX-92.

Ключевые слова: цифровая обработка сигналов, обнаружение речевой активности, гармоничность, модуляционная частота, отношение сигнал/шум

Для цитирования: Нефедов Н. Н., Алимурадов А. К. Краткий обзор способов обнаружения речевой активности // Инжиниринг и технологии. 2024. Т. 9 (2). С. 1–6. doi: 10.21685/2587-7704-2024-9-2-9

A brief overview of speech activity detection methods

Nikita N. Nefedov
Penza State University, 40 Krasnaya Street, Penza, Russia
nikita.nefyodow@mail.ru

Alan K. Alimuradov
Penza State University, 40 Krasnaya Street, Penza, Russia
alansapfir@yandex.ru

Abstract. Methods is presented, which have shown success in various signal-to-noise ratio (SNR) environments. VAD is the process of identifying speech segments in a continuous audio stream and is used to reduce computations and minimize potential recognition errors. The article discusses methods based on harmonicity and modulation frequency, which have demonstrated good performance in highly noisy environments. The article also presents the results of experiments evaluating harmonicity and modulation frequency based on VAD in a complex task involving remote speech detection for autonomous forklift control. It was found that the combination of harmonicity and modulation frequency can reduce false alarms and approach the accuracy of VAD systems where the boundary between speech and pause is manually designated. Further, an overview of the VAD system based on SUB-BAND is presented, which utilizes the maximum SNR values of sub-bands as detection functions. Experimental results show that the proposed method achieves better performance compared to the conventional ETSI AMR VAD in the NOISEX-92 database.

Keywords: digital signal processing, speech activity detection, harmonicity, modulation frequency, signal-to-noise ratio

For citation: Nefedov N.N., Alimuradov A.K. A brief overview of speech activity detection methods. *Inzhiniring i tekhnologii = Engineering and Technology*. 2024;9(2):1–6. (In Russ.). doi: 10.21685/2587-7704-2024-9-2-9

Обнаружение речевой активности (Voice Activity Detection, VAD) – это процесс идентификации сегментов речи в непрерывном аудиопотоке. VAD часто является первым этапом приложений для обработки речи и используется как для сокращения вычислений за счет исключения ненужной передачи и обработки неречевых сегментов, так и для уменьшения потенциальных ошибок распознавания в таких сегментах. Поскольку процесс принятия бинарных решений о наличии или отсутствии речи чреват ошибками, VAD часто избегают путем повторного обращения к пользователю с просьбой инициировать запись речи (например, с помощью механизмов push-to-talk или tap-and-talk).

Алгоритмы и программы VAD привлекают значительное внимание исследовательского сообщества. В условиях высокого качества записи методы, в основе которых лежит анализ энергетических параметров сегмента речи, показывают хорошие результаты [1]. Однако в условиях сильно зашумленной обстановки такие методы часто дают значительное число ложных срабатываний (ошибок первого рода). По этой причине изучаются и альтернативные методы, более выгодные для использования при малом отношении сигнал/шум (Signal-to-Noise Ration, SNR). Но для их эффективности требуется настройка параметров под конкретную шумовую среду, что создает трудности при работе с нестационарными и мгновенными типами шумов, которые встречаются достаточно часто в области цифровой обработки сигналов и VAD, в частности.

В данной статье представлен обзор альтернативных методов VAD, которые успешно себя продемонстрировали в ходе практического применения в средах с различными значениями.

В ситуациях с низкими значениями SNR несонорные части речевого сигнала обычно первыми становятся неслышимыми и маскируются шумом. Напротив, гармоники, связанные с основными формантами в вокализованных областях, имеют наилучшее значение SNR и наиболее устойчивы к аддитивной помехе (шуму). Поэтому был разработан детектор голосовой активности с использованием двухэтапного подхода, основанного на двух отличительных характеристиках речи – гармоничности и частоты модуляции (Modulation Frequency, MF). Модифицированная метрика гармоничности используется в качестве управляющей функции для набора параллельных классификаторов, включающих MF, рассчитанные в различных частотных диапазонах.

С помощью преобразования Фурье (конкретно в этой работе было применено кратковременное преобразование Фурье – STFT) выделяются гармонические компоненты, а далее – гармонические частоты, путем поиска пиковых значений в частотном спектре аудиосигнала. Для более точного определения голосовой активности вкупе с гармоничностью используются MF, отражающие изменения в частоте основного тона. Таким образом, высокая гармоничность и низкая частота модуляции могут указывать на наличие речи на конкретном интервале времени исследуемого сигнала.

Для периодического сигнала наибольший локальный максимум будет приходиться на время запаздывания τ, соответствующее периоду. Относительная мощность между локальным максимумом и пиком с нулевым запаздыванием соответствует количеству периодичности в сигнале. Гармоничность определяется как:

$$H = 10\log_{10}\frac{r_x(\tau)}{r_x(0)-r_x(\tau)},$$

где $r_x(\tau)$ – нормированная автокорреляция стационарного сигнала $x(t)$, τ – время запаздывания:

$$r_x(\tau)=\frac{\int x(t)\omega(t)x(t+\tau)\omega(t+\tau)dt}{\int \omega(t)\omega(t+\tau)dt},$$

где $\omega(t)$ – оконная функция (в данном случае – окно Хэннинга).

Методы контролируемого обучения (Neighborhood Components Analysis, NCA) и опорных векторов (Support Vector Machine, SVM) используются для классификации аудиопотоков на основе гармонической и MF, т.е. «звеном» принятия решения «речь» или «пауза» является модель, построенная на основе этих двух методов.

Упрощенная блок-схема алгоритма работы такой системы VAD представлена на рис. 1.

Рис. 1. Упрощенная блок-схема системы VAD

Для проверки работоспособности этого алгоритма была проведена серия экспериментов по оценке гармоничности и MF на основе VAD на сложной задаче, включающей обнаружение удаленной речи (команд) для управления автономным погрузчиком.

В этой части эксперимента начали с исследования гармоничности и MF в классификации речь/пауза как чистой задачи обнаружения, т.е. для оценки потенциала систем в различении речи и паузы.

Был сформирован синтетический набор данных, состоящий из речевых команд от 26 испытуемых с добавлением шума для имитации различных значений SNR в диапазоне от –5 до 15 дБ. И речь, и шум были реальными данными, записанными с помощью микрофона. Шумовые данные состояли из различных записей, с которыми обычно сталкиваются в среде работы автономных погрузчиков, включая шум двигателя, улицы, погрузочной платформы, фоновые разговоры и звуки окружающей среды, такие как ветер и т.д. Классификация проводилась по каждому фрагменту без каких-либо дополнительных шагов постобработки. Фрагменты с переходом между речью и паузой исключались.

Для сравнения были включены результаты систем, оптимизированных для этой задачи на основе относительной спектральной энтропии (Relative Spectral Entropy, RSE) [2], долгосрочной изменчивости сигнала (Long-Term Signal Variability, LTSV) [3], и VAD на основе статистических моделей (таких как в работе [4]), использующих мел-частотные кепстральные коэффициенты (Mel-Frequency Cepstral Coefficients, MFCC) и MF в качестве признаков для моделей гауссовой смеси (Gaussian Mixture Model, GMM).

Результаты работы различных конфигураций VAD на синтетических данных приведены в табл. 1. Способ на основе GMM и MFCC показал самые высокие значения коэффициента равной вероятности ошибок первого и второго родов (Equal Error Rates, EER) и значений ошибки второго рода (False Acceptance Rate, FAR). В то время как способ на основе GMM и MF оказался достаточно конкурентоспособным. Это может быть связано с тем, что MFCC очень восприимчивы к шуму. Система VAD на основе RSE обеспечила лучшие результаты, чем система на основе гармоничности. Однако комбинирование систем на основе MF и гармоничности позволило снизить значение FAR максимум на 71 % по сравнению с системой RSE.

Таблица 1

Сравнительные показатели различных методов

Система VAD	EER, %	FAR, %
RSE	7,92	18,21
LTSV	20,1	71,75
GMM:MFCC	31,75	61,12
GMM:MF	7,4	19,5
Harmonicity	12,93	56,69
SVM:MF_{whole}	5,08	14,23
SVM:MF_{ch}	4,53	8,99
SVM:MF_{whole} + NCA	4,93	11,47
SVM:MF_{ch} + NCA	4,15	8,38
SVM:MF_{ch} + NCA + Harm	4,01	7,65

Затем были рассмотрены различные конфигурации MF. Одиночная конфигурация SVM (MF_{whole}) (основанная на работе [5]) показала результаты хуже, чем параллельная конфигурация SVM (MF_{ch}), которая более устойчива к искажениям сигнала от шумов, ограниченных по частоте. А использование NCA-преобразований перед способом на основе SVM улучшило производительность в обоих случаях. Наилучшие показатели VAD были получены, когда гармоничность была добавлена в качестве функции стробирования к параллельному способу на основе SVM, что позволило снизить FAR на 9 % по сравнению с лучшей системой, основанной исключительно на MF.

Таким образом, комбинирование гармоничности и MF помогла снизить FAR еще на 9 % в обстановке с нестационарным шумом и по точности приблизился к VAD, в которых границу между речью и паузой сегментируют вручную.

Известно, что энергия речевого сигнала неравномерно распределена по всем частотам. Например, энергия вокализованной речи в основном концентрируется в областях формант. Это означает, что локальные значения SNR в формантных областях становятся значительно выше по сравнению с другими областями. Когда чистая речь зашумляется сильным фоновым шумом, некоторые частотные биты речевого кадра все еще могут иметь высокие значения SNR (указанные стрелками) [6]. С другой стороны, когда речь отсутствует, значения SNR всех частотных битов, как правило, равномерно низкие. Таким образом, лучше использовать SNR-ориентированные функции, чем функции, основанные на энергии, для представления речевого сигнала. Во многих случаях, поскольку шумовой сигнал содержит низкочастотные компоненты и маскирует содержание речи, использование только глобального значения SNR может не работать. Поэтому осуществляется извлечение функции, основанной на локальных значениях SNR, чтобы уменьшить шум.

Учитывая спектр речевого высказывания, преобразованный с помощью (Discrete Fourier Transform, DFT), изначально делится весь спектр на несколько поддиапазонов. Во-вторых, оцениваются значения SNR поддиапазонов, и максимальные значения SNR поддиапазонов (Maximum Values of Sub-Band SNR, MVSS) извлекаются в качестве детекторных функций. Оценка фона и окончательное решение VAD выполняются путем сравнения значения функции с оцененным порогом. При инициализации оценивается спектр шума и порог, предполагая, что речь всегда следует за начальным периодом шума.

Адаптивный порог применяется для повышения точности VAD и быстрого отслеживания зашумленного сигнала без сложных вычислений. Экспериментальные результаты показывают, что предложенный метод достигает лучших показателей по сравнению с обычными European Telecommunications Standards Institute, Adaptive Multi-Rate (ETSI AMR) VAD в базе данных NOISEX-92.

Экспериментальные результаты демонстрируют потенциальное преимущество использования MVSS в качестве детекторной функции. На рис. 2 представлено сравнение различных характеристик в низкочастотном диапазоне. Из рисунка видно, что график энергии медленно меняется вблизи границы между речью и паузой, что затрудняет обнаружение речи. Напротив, график MVSS значительно меняется между речью и паузой, что облегчает обнаружение.

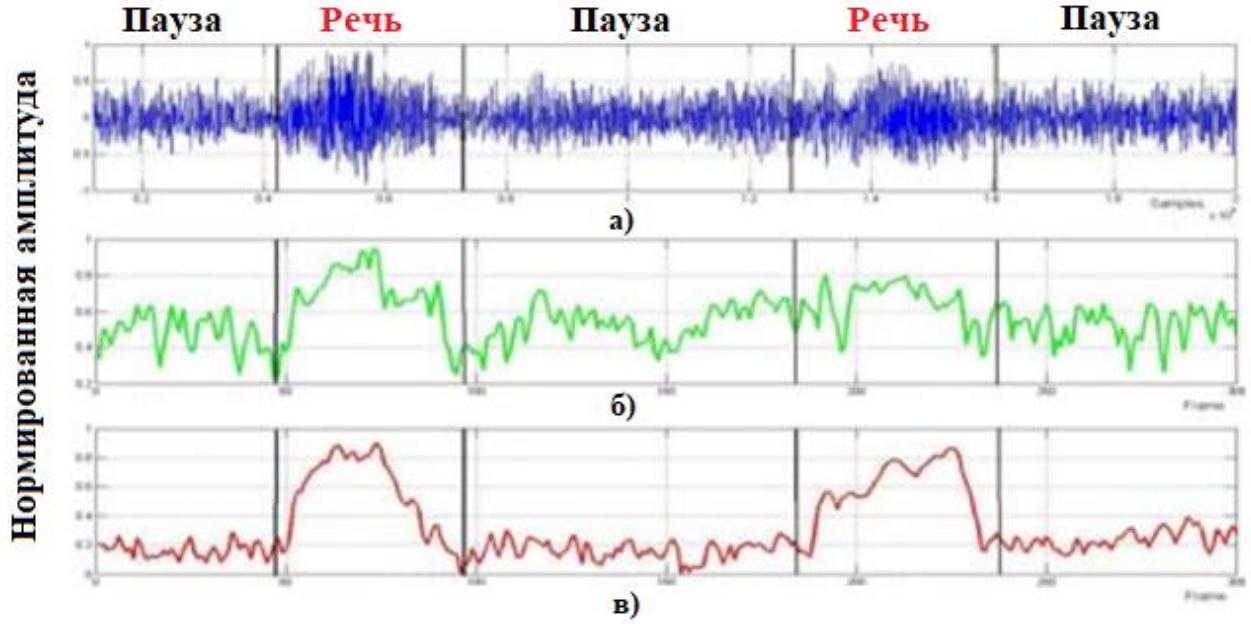

Рис. 2. Сравнение результатов сегментации: *а* – нормированная амплитуда зашумленного мандаринского произнесения «три, четыре»; *б* – сумма энергии субполос (0~1000 Гц); *в* – сумма MVSS (0~1000 Гц)

Тестовая база данных, используемая для исследований, была собрана из 20 отдельных дикторов (10 мужчин и 10 женщин). Каждый диктор в спокойной обстановке прочитал 5 фрагментов 10 цифр

на китайском языке «1, 2, ..., 10», и речь была записана. Использовались различные шумы из базы данных NOISEX-92 [7], включая белый шум, розовый шум, шум автомобиля и шум военной техники. Входной сигнал был оцифрован с частотой 8000 Гц. Затем были выделены фрагменты длительностью 32 мс (с перекрытием 24 мс) и применена оконная функция Хэмминга. Наконец, было применено 256-точечное быстрое преобразование Фурье (БПФ). Предложенный способ VAD оценивается с точки зрения его способности идентифицировать сегменты речи или паузы (фоновый шум) при различных значениях SNR. Решения об идентификации чистой речи принимались путем ручной маркировки образцов. Эффективность обнаружения оценивается с точки зрения частоты попаданий в речь (Speech Hit Rate, SHR) и частоты попаданий в неречевые фрагменты – паузы (Non-Speech Hit Rate, NSHR).

Сравнение эффективности различных методов проводится с точки зрения SHR и NSHR в диапазоне от 15 дБ до 0 дБ. Отмечается, что способ на основе AMR VAD1 стабильно работает с высокими значениями SHR во всем диапазоне значений SNR. Однако при увеличении уровня шума способ демонстрирует низкую эффективность с точки зрения NSHR.

VAD2 демонстрирует значительные улучшения по сравнению с VAD1, усовершенствуя значение NSHR. К сожалению, по-прежнему быстро ухудшается эффективность обнаружения речи при определенных неблагоприятных условиях зашумления (например, розовый шум). Предложенный способ может справиться с вышеуказанными проблемами даже при снижении значения SNR и достигает наилучшей производительности в среднем по показателям SHR и NSHR.

В настоящее время VAD системы на основе нейронных сетей находят все большее распространение ввиду своей эффективности в любых условиях. Однако, несмотря на значительные достижения в области VAD на базе нейронных сетей, существуют вызовы и проблемы, требующие дальнейших исследований. Например, улучшение производительности в условиях высокого уровня шума, оптимизация работы систем при ограниченных вычислительных ресурсах, а также повышение устойчивости к различным типам помех.

Альтернативные VAD могут являться надежной заменой быстро развивающимся системам с искусственным интеллектом. VAD, представленные в данной обзорной статье, обладают значительной эффективностью при достаточно простом алгоритме работы.

Список литературы

1. Chuangsuwanich E., Glass J. Robust Voice Activity Detector for Real World Applications Using Harmonicity and Modulation Frequency // Interspeech. 2011. 28–31 August. P. 2645–2648.
2. Ouzounov A. Robust features for speech detection – a comparative study // International Conference on Computer Systems and Technologies. 2005. P. 19/1–19/6.
3. Ghosh P., Tsiartas A., Narayanan S. Robust voice acitivity detection using long-term signal variability // IEE Trans. Audio, Speech and Language Processing. 2011. № 19. P. 600–613.
4. Sohn J., Kim N. and Sung W. A statistical model-based voice activity detection // IEEE Signal Processing Letters. 1999. № 6. P. 1–3.
5. Bach J., Kollmeier B., Anemuller J. Modulation-based detection of speech in real background noise: Generalization to novel background classes // ICASSP. 2010. P. 41–44.
6. Jiang W., Lo W. K., Meng H. A new voice activity detection method using maximized Sub-band SNR // International Conference on Audio, Language and Image Processing, Shanghai. China; 2010. C. 80–84.
7. Varga A. P., Steeneken H. J. M., Tomlinson M., Jones D. The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition // Technical Report, DRA Speech Research Unit. 1992.

References

1. Chuangsuwanich E., Glass J. Robust Voice Activity Detector for Real World Applications Using Harmonicity and Modulation Frequency. *Interspeech*. 2011;28–31 August:2645–2648.
2. Ouzounov A. Robust features for speech detection – a comparative study. *International Conference on Computer Systems and Technologies*. 2005:19/1–19/6.
3. Ghosh P., Tsiartas A., Narayanan S. Robust voice acitivity detection using long-term signal variability. *IEE Trans. Audio, Speech and Language Processing*. 2011;(19):600–613.
4. Sohn J., Kim N. and Sung W. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*. 1999;(6):1–3.
5. Bach J., Kollmeier B., Anemuller J. Modulation-based detection of speech in real background noise: Generalization to novel background classes. *ICASSP*. 2010:41–44.

6. Jiang W., Lo W. K., Meng H. A new voice activity detection method using maximized Sub-band SNR. *International Conference on Audio, Language and Image Processing, Shanghai*. China, 2010:80–84.
7. Varga A.P., Steeneken H.J.M., Tomlinson M., Jones D. The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. *Technical Report, DRA Speech Research Unit*. 1992.

Поступила в редакцию / Received 13.03.2024

Принята к публикации / Accepted 13.04.2024