E

УДК 004.89,004.93

Т. В. Ермоленко, А. П. Тихончук
ГОУ ВПО «Донецкий национальный университет (ДОННУ)»
83001, г. Донецк, ул. Университетская, 24

ОПРЕДЕЛЕНИЕ ГОЛОСОВОЙ АКТИВНОСТИ В РЕЧИ

T. V. Ermolenko, A. P. Tihonchuk
SEI HPE «Donetsk National Technical University(DONNU)»
83001, c. Donetsk, st. Universitetskay 24

VOICE ACTIVITY DETECTION IN SPEECH

Т. В. Ермоленко, А. П. Тихончук
ДНЗ ВПО «Донецький національний університет(ДОННУ)»
83001, м. Донецьк, вул. Університетська, 24

ВИЗНАЧЕННЯ ГОЛОСОВОЇ АКТИВНОСТІ У МОВЛЕННІ

В статье рассматривается задача определения границ речи в звуковом сигнале. Выполнен анализ исследований в области обработки речи. На основе результатов проведено сравнение алгоритмов выделения голоса, в результате которого сделан вывод, что наличие речи увеличивает количество информации в соответствующих местах звукозаписи. Предложен и реализован алгоритм, использующий информационную энтропию для классификации частей звукозаписи по признаку наличия/отсутствия речи в присутствии произвольных помех. Выполнены численные исследования эффективности алгоритма.
Ключевые слова: энтропия, распознавание речи, алгоритм распознавания.

The article deals with the problem of determining the boundaries of speech in a sound signal. The analysis of researches in the field of speech processing is carried out. Based on the results, a comparison is made between the algorithms of voice allocation, this concluded that the presence of speech increases the amount of information in the corresponding places of sound recording. An algorithm that uses information entropy for the classification of parts of sound recording based on the presence/absence of speech in the presence of arbitrary interference is proposed and implemented. Numerical studies of the efficiency of the algorithm are performed.
Keywords: entropy, speech recognition, recognition algorithm.

У статті розглядається задача визначення границь мовлення у звуковому сигналі. Виконаний аналіз досліджень у сфері обробки мовлення. На основі результатів проведено порівняння алгоритмів виділення голосу, в результаті якого дійшли висновку, що наявність мовлення збільшує кількість інформації у відповідних місцях звукозапису. Запропоновано і реалізовано алгоритм, який використовує інформаційну ентропію для класифікації частин звукозапису за ознакою наявності/відсутності мови в присутності довільних перешкод. Виконано чисельні дослідження ефективності алгоритму.
Ключові слова: ентропія, розпізнавання мови, алгоритм розпізнавання.

Общая постановка проблемы

Шум является существенной проблемой для систем обработки звукового сигнала, в которых решается задача определения активности речи, поиск отдельных фраз, слов или звуков. Системам распознавания голоса, улучшения качества звука, сжатия речевого сигнала и т.д. необходимо взаимодействовать с самыми разными устройствами с разными техническими характеристиками в разном звуковом окружении. Известным алгоритмам определения голоса и подавления помех зачастую требуется обучение. Для обучения используются образцы шума – не содержащие речь фрагменты звука.

В условиях, когда помехи непредсказуемы, а признаки шума и голоса заранее неизвестны, дальнейшая обработка речи затруднительна. Поэтому задача классификации отдельных частей звукового сигнала по признаку наличия или отсутствия речи при наличии заранее неизвестных, динамически изменяющихся помех, является крайне актуальной. Задача определения активности речи и классификации звукового сигнала, с учетом упомянутых выше условий, совсем не тривиальна. Для большинства известных алгоритмов детектирования голосовой активности свойственно заметное, а иногда критичное, увеличение ошибок в тех случаях, когда меняется характер шума или увеличивается его уровень.

Анализ исследований и публикаций. Входной сигнал представляет собой последовательность отсчетов $<x_0, x_1, \ldots, x_N>$ – значений сигнала, взятых через определенные промежутки времени (с частотой дискретизации). Входной сигнал делится на фреймы – участки постоянной длины. Промежуток времени, соответствующий каждому фрейму, должен быть достаточно мал для того, чтобы в его пределах сигнал оставался относительно постоянным и, в то же время, содержать достаточно отсчетов для дальнейшей классификации фрейма по рассчитанным для него признакам. Например, типичным значением для систем распознавания речи являются фреймы длиной 20–25 мс, расположенные с шагом 10–12,5 мс [1].

Типичный алгоритм детектирования голосовой активности на каждом фрейме последовательно выполняет: извлечение признаков, классификацию фрейма на наличие или отсутствие в нем речи и сглаживание результатов классификации (рис. 1).

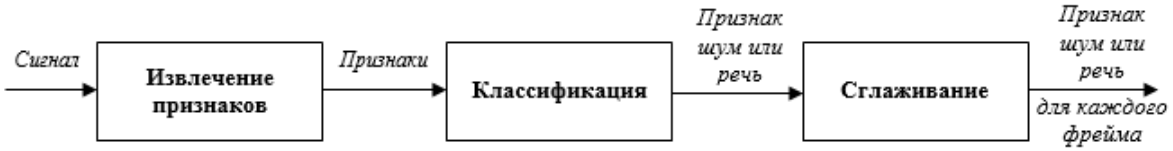

Рисунок 1 – Схема типичного алгоритма детектирования голоса

Модуль извлечения признаков получает числовые характеристики из звукового сигнала для каждого фрейма и формирует последовательность фреймов с вычисленными признаками с целью проведения дальнейшей классификации.

Модуль классификации получает признаки, а возвращает имя класса каждого фрейма: шум или речь. В зависимости от того, какие признаки используются, реализация модуля выполняется одним из методов машинного обучения [2]: метод опорных векторов, нейронные сети различной архитектуры, линейные, вероятностные классификаторы и т.д.

Модуль сглаживания предназначен для уточнения результатов классификации исходя из стандартной длительности фонем. Так, единственный фрейм длиной 20–25 мс

в окружении шума не может быть речью, потому как у человека самые короткие фонемы занимают больший промежуток времени. Также не приспособлен звуковой аппарат к миллисекундным паузам в середине слова.

Основным способом улучшения алгоритмов детектирования речи является выбор таких признаков, которые бы позволили достоверно отличать речь от шума. Энергия сигнала – первый практически используемый признак [3]. Это простой признак, который может быть использован только в случае умеренного шума, когда соотношение сигнал/шум превышает 30 дБ. Основной недостаток данного признака – необходимость предварительного задания порогового значения. Усовершенствованными алгоритмами детектирования на основе энергии являются: рекурсивная оценка шума [4], использование гистограмм или огибающих [5], учёт энергии нескольких последовательных фреймов [6], сравнение с найденным эталонным фреймом [7]. Как показало сравнение [8], выделение фрейма с шумом и его дальнейшее сопоставление с другими фреймами [7], показывает наилучшие результаты.

Другой способ улучшения основывается на тональности звука. В соответствии с моделью произношения звуков [9], речь моделируется звонкими и глухими возбуждающими сигналами, которые потом модулируются речевым аппаратом. Для гласных и звонких согласных звуков голосовые связки порождают гармонический голосовой тон частоты от 50 до 250 Гц, что позволяет найти их в звуковом сигнале. Однако глухие и шипящие звуки определить таким способом сложно, более того, музыка тоже часто определяется как речь [10]. К признакам, использующим тональность звука, относятся частота изменения знака сигнала [3], нормированная автокорреляционная функция, спектральная энтропия, размах значений кепстральных компонент [10], комбинация частоты смены знака с энергией [11], логарифм произведения подмножества спектральных компонент [12]. Лучшим среди признаков, опирающихся на анализ тональности [8], является разность максимального и минимального значений кепстральных компонент, который позволяет достичь 80–85% точности даже при опознавания глухих звуков.

Дополнительной к основной тональности, признаком речи является её форманта – акустическая характеристика звуков речи (прежде всего гласных), связанная с уровнем частоты голосового тона и образующая тембр звука. В идеальном случае форманта описывается формой спектра, который является вектором потенциально бесконечной размерности. Однако даже кепстральные коэффициенты низкого порядка [13], мелчастотные кепстральные коэффициенты (MFCC) [14], коэффициенты линейного предиктивного кодирования [15] позволяют достичь приемлемых результатов. Для определения речи на основании формы спектра, многомерные векторы признаков группируют, заранее составляя справочник векторов с помощью машинного обучения. Общий недостаток указанных алгоритмов – необходимость обучения классификатора, что приводит к возрастанию числа ошибок в условиях непредсказуемого шума.

Обычно речь изменяется заметно быстрее, чем шум. Для вычисления степени стационарности в работе [16] рассматривались промежутки времени большей длины, чем обычная продолжительность фонемы, и на основе спектра нескольких фреймов определялась величина долговременной вариации сигнала. Недостаток подхода – значительное число ошибок в условиях нестационарного шума.

Признаки, основывающиеся на свойстве ритмичности чередования согласных и гласных звуков в человеческой речи, устойчивы к помехам. Однако для их вычисления необходимо рассматривать промежутки времени около 1 секунды. Например, спектрально-темпоральная модуляция [17] рассматривает модуляцию с изменением

одновременно времени и частоты, стремясь смоделировать восприятие звука человеком, а также учитывает тональную и форматную структуру речи [18]. К недостаткам можно отнести вектор признаков большой размерности – в некоторых случаях больше 1000, - из-за чего классификатор приходится обучать на большом количестве примеров, кроме того, обработка коротких звукозаписей затруднена из-за длинных фреймов.

Статистический подход [19], [20] основывается на предположении, что речь и шум имеют разные спектры, каждую компоненту которых можно описать неким распределением вероятности, и, вычислив отношение функций правдоподобия, получить статистический классификатор, разделяющий шум и человеческую речь.

Кроме единичных признаков, исследовался вопрос комбинации признаков [21].

Использование энтропии для детектирования голосовой активности

Результаты проведенного сравнения алгоритмов выделения голоса однозначно говорят о том, что речь является носителем информации. Тогда естественно предположить, что добавление речевого сигнала к фоновому шуму увеличит количество информации в соответствующих местах звукового сигнала, а значит, даёт возможность использовать рассчитанное значение информационной энтропии для определения речи в сигнале.

Энтропия – это мера беспорядка, мера неопределённости какого-либо опыта. В случае речевого сигнала энтропия характеризует степень его нестационарности.

Для того, чтобы подсчитать энтропию сигнала на фрейме необходимо выполнить следующие действия:

– нормализовать входящий сигнал таким образом, чтобы его значения лежали в диапазоне [-1;1];

– построить гистограмму (плотность распределения) значений сигнала во фрейме. Фрейм делится на L ($L < K$) частей $[a_0, a_1], [a_1, a_2], \dots, [a_{L-1}, a_L]$, где $a_0 = x_k$ и $a_L = x_k + K$, для каждой части подсчитывается количество амплитуд – составляется гистограмма частот;

– информационная двоичная энтропия для независимых случайных событий x с N возможными состояниями, распределённых с вероятностями p_k ($k=0,\dots,N-1$), рассчитывается по формуле:

$$H_f = -\sum_{k=0}^{N-1} p[k] \bullet \log(p[k]), \quad (1)$$

где $p(k) = \frac{|x_k|}{\sum_{m=0}^{n-1} |x_m|}$ – спектральная плотность k-й компоненты спектра, x_k – спектральные коэффициенты.

Полученное значение энтропии всего лишь ещё одна характеристика фрейма, и для того, чтобы отделить речь от шума, по-прежнему необходимо её с чем-то сравнивать.

Одним из вариантов использования энтропии является использование порога, равного среднему между её максимальным и минимальным значениями (среди всех фреймов). Однако, как показал опыт, такой подход не дал сколь либо хороших результатов.

Энтропия (в отличие от того же среднего квадрата значений амплитуды сигнала) – величина робастная (устойчивая к шумам, искажениям и уровню записи),

что дает возможность подобрать значение её порога в виде константы. Такой подход показал заметно лучший результат, чем среднее. Однако процент ошибок всё еще значителен, в особенности при изменении характера и уровня шума, при том же значении порога. И только после незначительной корректировки порогового значения, результаты стабилизируются.

Для повышения качества детектирования речи может быть использована комбинация нескольких признаков [21]. Была проверена комбинация энтропии с энергией сигнала и MFCC [22]. При этом сравнение энтропии с порогом служит для вычисления порога второго признака. Такой подход позволяет корректировать значение порога для энтропии, анализируя разницу результатов классификации и сглаживания.

Результаты практического эксперимента показали, что число ошибок детектирования меньше при использовании более простого признака энергия.

Энергия вычисляется как смещённая оценка дисперсии входного сигнала:

$$E = \frac{1}{N}\sum_{k=0}^{N-1}([x_k - \bar{x})]^2, \qquad (2)$$

где $\bar{x}$ – среднее значение сигнала, N – количество отсчетов сигнала во фрейме.

Алгоритм детектирования речи, использующий энтропию

Общая структура алгоритма, использующего энтропию, включает в себя обучение системы, классификацию фреймов, вычитание спектра шума (рис. 2).

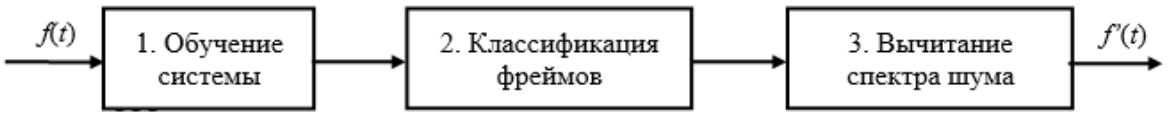

Рисунок 2 – Схема работы модели

На блок «Обучение системы» поступает входящий сигнал *f(t)*. Предполагается, что в начале, на заданном временном участке (по умолчанию его продолжительность равна 1 сек.), сигнал содержит только шум. В это время выполняется первоначальный расчет порога для вторичного признака (энергии). Если уровень шума незначителен, обучение занимает меньше времени, классификация может быть начата раньше.

Для каждого *s*-го фрейма хранится информация:

– рассчитанные значения энтропии и энергии. Данные значения используются в процессе обучения, а также для классификации фреймов типа «пауза»;

– признак того, что значение энтропии больше или меньше порога. Необходимость хранить этот признак связана с тем, что значение для порога энтропии меняется в процессе обучения и работы алгоритма;

– значение порога классификации фрейма. Необходимость хранить этот параметр связана с тем, что значение порога, на основании которого выполнялась классификация, меняется в процессе обучения и работы алгоритма;

– признак того, что фрейм был определен как шум, речь, пауза или не определен.

Фрейм считается неопределенным, если нельзя достоверно отнести его к шуму или речи по причине нехватки собранной статистической информации (в частности, в начале процесса получения сигнала, пока не получен фрейм, значение энтропии которого ниже порога и/или не получен порог для энергии).

Значения энтропии уменьшаются на гласных и могут резко возрастать из-за шумов, присутствующих в щелевых и смычно-щелевых звуках. Для того чтобы бороться с первой проблемой, приходится вводить понятие «минимального расстояния между словами», вторая проблема решается использованием «минимальной длины слова», что позволит уточнить результаты определения голоса на основании типичной длительности фонем. Фрейм, определенный как шум и следующий за фреймом, определенным как речь, может на самом деле являться частью речи. Такой фрейм помечается как пауза, затем, в процессе анализа последующих и предыдущих фреймов такой фрейм будет отнесен или к шуму, или к речи.

В работе алгоритма используются фреймы за заданное время (последние 10 с). В связи с этим, с одной стороны, алгоритм обладает свойством самообучения, с другой – мало зависит от сигнала, который был достаточно давно.

Входными данными алгоритма являются

отсчеты звукового сигнала $\{x_k\}$, k=0,…, N-1;

L_{min} – минимальная длина фонемы (в фреймах).

L_{max} – максимальная длина паузы внутри слова (в фреймах).

Выходные данные алгоритма:

массив маркировки фреймов $Mark = \{Mark(s)\}$, s=1,…, N_{fr},

где N_{fr} – число фреймов входного сигнала,

$$Mark(s) = \begin{cases} 0, \text{ åñëè ô ðåéì ñî äåðæèò òî ëüêî ø óì} \\ 1, \text{ åñëè ô ðåéì ñî äåðæèò òî ëüêî ðå÷ü} \\ 2, \text{ åñëè ô ðåéì ñî äåðæèò ï àóçó â ãëóõèõ ñì û÷í û õ (ï ,ê,ò)} \\ -1, \text{ åñëè êëàññ ô ðåéì à í å î ï ðåäåëåí} \end{cases}$$

L, R – номера фреймов, являющиеся левой и правой границами слова соответственно.

0. Инициализация:

0.1. Если длина входного сигнала менее секунды, то останов.

0.2. Иначе

0.2.1. Формируем множество *Noise* – множество номеров фреймов, содержащих только шум. Полагаем, что в первую секунду записи речи нет, поэтому изначально номера фреймов первой секунды включаются в это множество. Для этих фреймов $Mark(s)$=0.

0.2.2. Формируем множество – множество номеров фреймов, содержащих только речь.

0.2.3. Вычисляем среднее и смещенную оценку дисперсии значений энтропии (1), а также значений энергии (2) на фреймах, содержащих только шум:

$$M_H = \frac{1}{|Noise|}\sum_{S\in Noise} H_f(s),\ D_H \frac{1}{|Noise|}\sum_{S\in Noise}(H_f(s)-M_H)^2 \quad ,M_E = \frac{1}{|Noise|}E(s), \quad (3)$$

где s – номер фрейма.

0.2.4. На основе полученных статистик вычисляем пороги:

$$h = M_H + \sqrt{D_H}\ , e = M_E \quad (4)$$

0.2.5. Для остальных фреймов полагаем $Mark(s)$=-1.

0.2.6. $L = R = 0$,

0.2.7. Обнуляем счетчик количества подряд идущих фреймов, содержащих речь: $T = 0$.

Для каждого последующего фрейма m:

1. Вычисляем согласно (1) значение энтропии $H_f(m)$ и энергии $E(m)$ согласно (2).

2. Если $H_f(m) < h$

2.1. То

2.1.1. Добавляем m во множество *Noise* и пересчитываем пороги (4).

2.1.2. $Mark(m)=0$.

2.1.3. $m = m+1$, переход на шаг 1.

2.2. Иначе

2.2.1. Добавляем m во множество $\overline{Noise}$.

2.2.2. Пересчитываем $e = \frac{1}{|Noise|} \sum_{s \in Noise} E(s)$

2.2.3. $m = m+1$, переход на шаг 1.

2.3. Если $E(m)>e$, то

2.3.1. Классифицируем фрейм как речь: $Mark(m)=1$.

2.3.2. Наращиваем счетчик количества подряд идущих фреймов, содержащих речь: $T = T+1$.

2.3.4. Если количество подряд идущих фреймов, содержащих речь, првышает минимальную длину фонемы $T \geq L_{min}$, то определяем левую границу речи $L = m - L_{min}$.

2.3.5. $m = m+1$, переход на шаг 1.

2.4. Иначе

2.4.1. Если не было зафиксировано начало речи ($L = 0$), то

2.4.1.1. $Mark(m)=0$ (классифицируем фрейм как шум).

2.4.1.2. Добавляем m во множество *Noise* и пересчитываем пороги (4).

2.4.1.3. Обнуляем счетчик количества подряд идущих фреймов, содержащих речь: $T = 0$.

2.4.1.4. $m = m+1$, переход на шаг 1.

2.4.2. Иначе (начало речи определено, $L > 0$)

2.4.2.1. Классифицируем фрейм как паузу, связанную с глухой смычной: $Mark(m)=2$.

2.4.2.2. Наращиваем счетчик подряд идущих фреймов, содержащих паузу: $P = P+1$.

2.4.2.3. Если $P \geq L_{max}$, то определяем правую границу речи $R = m - L_{max}$.

2.4.2.4. $m = m+1$, переход на шаг 1.

После определения границ речи на фреймах, содержащих шум, вычисляются усредненные спектральные характеристики шума, затем на каждом фрейме, содержащем речь, выполняется спектральное вычитание – от спектра сигнала вычитается спектр шума, по полученным спектральным коэффициентам сигнал восстанавливается.

Анализ эффективности программной реализации алгоритма

В качестве критерия эффективности алгоритма выступает вероятность верного определения последовательности фреймов, являющихся непрерывным фрагментом речи, ограниченных шумом.

Для анализа эффективности распознавания речи реализованным алгоритмом в зависимости от отношения сигнал/шум (SNR) использовалось специализированное ПО. Задача генерации шума была решена с помощью звукового редактора CoolEdit. Для измерения отношения сигнал/шум использовался анализатор спектра SpectraLAB.

Была проведена серия экспериментов. Условия проведения экспериментов следующие:

1. Количество дикторов с разными голосовыми характеристиками: 3;
2. Количество различных слов: 32;
3. Значения SNR: 60дБ, 18дБ, 12дБ, 6дБ;
4. Число повторений каждого слова одним диктором: 10;
5. Для каждого эксперимента фиксировалась следующая информация:

– Определение границ речи: Да / Нет

– Точность определения левой и правой границ речи в % согласно экспертной оценке диктора;

Тестировались только заданные слова, другая речь не тестировалась.

В ходе каждого эксперимента, тестеры заполняли форму (рис. 3).

Слово	SNR = 60дБ	№ эксперимента			
		1	2	...	10
New	Речь, да/нет				
	Левая граница, %				
	Правая граница, %				
Open	Речь, да/нет				
	Левая граница, %				
	Правая граница, %				
...					
Right	Речь, да/нет				
	Левая граница, %				
	Правая граница, %				

Рисунок 3 – Заполняемая в процессе тестирования форма

В результате анализа всех полученных результатов сформированы усредненные данные, которые отражены в результирующей табл/1.

Таблица 1 – Исследование корректности и работоспособности алгоритма в зависимости от SNR

SNR	Процент ошибок	Точность определения границ речи	
		Левая граница	Правая граница
60дБ	2%	+1%	+3%
18дБ	7%	-2%	+4%
12дБ	20%	+5%	-3%
6дБ	55%	-10%	+12%

Выводы

В работе предложен и реализован на практике способ использования информационной энтропии в алгоритме детектирования речи. Очевидными преимуществами предложенного алгоритма детектирования речи с применением энтропии являются:

1. Меньшее число ошибок в случае неоднородного, изменяющегося шума, за счет: а) динамического расчета порога для вторичных признаков на основании сравнения энтропии с порогом энтропии; б) корректировки порога для энтропии; в) устойчивости энтропии к шумам;
2. Возможность обработки звукового сигнала в реальном времени;
3. Малое время начального обучения. В случае отсутствия шума, детектирование речи начинается практически с самого начала работы алгоритма.

Выполненное исследование на практике показало, что представленный способ позволяет даже в случае сильной зашумленности обнаружить в звуковом сигнале человеческую речь. Разработанный алгоритм детектирования речи предполагается использовать в системах автоматической обработки звукового сигнала. Программная реализация алгоритма применена на практике в программном комплексе речевого интерфейса текстового процессора.

Список литературы

1. Jurafsky D. Speech and Language Processing (2nd Edition) [Text] / Jurafsky, Daniel and Martin, James H. – NJ. : Prentice Hall, 2009. – 1024 p.
2. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Текст] / Петер Флах. – М. : ДМК Пресс, 2015. – 400 с.
3. Rabiner L. R. An algorithm for determining the endpoints of isolated utterances [Text] / L. R. Rabiner, M. R. Sambur // Bell Syst. Tech. J. – 1975. – V. 54, № 2. – P. 297–315.
4. Van Gerven S. A comparative study of speech detection methods [Text] / S. Van Gerven, F. Xie // Proc. of European Conference on Speech, Communication and Technology. – Rhodos, 1997. – Режим доступа : http://www.mirlab.org/conference_papers/International_Conference/Eurospeech%201997/pdf/tab/a0199.pdf.
5. Marzinzik M. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics [Text] / M. Marzinzik, B. Kollmeier // IEEE Trans. Speech Audio Process. – 2002. – V. 10, № 2. – P. 109–118.
6. Ramírez J. Efficient voice activity detection algorithms using long-term speech information [Text] /J Ramírez, J. C. Segura, C. Benítez, Á de la Torre, A Rubio // Speech Commun. – 2004. – V. 42. – P. 271–287.
7. Pencak J. The NP speech activity detection algorithm [Text] / J. Pencak, D. Nelson // Proc. of ICASSP. – Detroit, 1995. – Режим доступа : https://www.researchgate.net/profile/Douglas_Nelson8/ publication/ 3618329_The_NP_ speech_activity_detection_algorithm/links/540857d90cf2c48563bb1228.pdf
8. Graf S. Features for voice activity detection: a comparative analysis [Text] / Simon Graf, Tobias Herbig, Markus Buck and Gerhard Schmidt // EURASIP Journal on Advances in Signal Processing . – 2015. – V. 2015, № 1. – P. 1–15.
9. Nelson D. J. Pitch-based methods for speech detection and automatic frequency recovery [Text] / D. J. Nelson, J. Pencak // Proc. of SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation. – San-Diego, 1995. –
   Режим доступа: https://www.researchgate.net/ profile/Douglas_Nelson8/publication / 260816047_Pitch-based_methods_for_speech_detection_and_automatic_frequency_recovery/links/ 541c15250cf2218008c4e563.pdf
10. Kristjansson T. Voicing features for robust speech detection [Text] / T. Kristjansson, S. Deligne, P. Olsen // Proc. of INTERSPEECH. – Lisbon. – 2005. – Режим доступа : http://papers.traustikristjansson.info/wp-content/ uploads/2011/07/KristjanssonRobustVoicingEurospeech2005.pdf
11. Shahnaz C. A multifeature voiced/unvoiced decision algorithm for noisy speech / C Shahnaz, W-P Zhu, MO Ahmad // Proc. of ISCAS. – Kos, 2006. – P. 2528–2531.
12. Sadjadi S. O. Unsupervised speech activity detection using voicing measures and perceptual spectral flux [Text] / S. O. Sadjadi, J. H. L. Hansen // IEEE Signal Process. Lett. – 2013. – V. 20, № 3. – P. 197–200.
13. Haigh J. A. A voice activity detector based on cepstral analysis [Text] / J. A. Haigh, J. S. Mason // Proc. of EUROSPEECH. – Berlin, 1993. – Режим доступа : https://www.semanticscholar.org/paper/A-voice-activity-detector-based-on-cepstral-HaighMason/0fc5b0a4d38a6ae1b5ce9bb347b82e3ef3505859/pdf
14. Kinnunen T. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data [Text] / T. Kinnunen, P. Rajan // Proc. of ICASSP. – Vancouver: IEEE, 2013 – P. 7229–7233.
15. Rabiner L. R. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem [Text] / L. R. Rabiner, M. R. Sambur // IEEE Trans. Acoust. Speech Signal Process. – 1977. – V. 25, № 4. – P. 338–343.
16. Ghosh P. K. Robust voice activity detection using long-term signal variability [Text] / P.K. Ghosh, A. Tsiartas, S. Narayanan // IEEE Trans. Audio, Speech, Lang. Process. – 2011. – V. 19, № 3. – P. 600–613.
17. Mesgarani N. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations [Text] / N. Mesgarani, M. Slaney, S. A. Shamma // IEEE Trans. Audio, Speech Lang. Process. – 2006. – V. 14, № 3. – P. 920–930.
18. Ezzat T. Spectro-temporal analysis of speech using 2-D Gabor filters [Text] / T. Ezzat, J. Bouvrie, T. Poggio // Proc. of INTERSPEECH. – Antwerp:ISCA, – 2007. – Режим доступа: http://cbcl.mit.edu/projects/cbcl/publications/ps/ezzat-spetro-analysis-07.pdf

19. Sohn J. A statistical model-based voice activity detection [Text] / J. Sohn, N. S. Kim, W. Sung // IEEE Signal Process. Lett. – 1999. – V. 6, № 1. – P. 1–3.
20. Chang J.-H. Voice activity detection based on multiple statistical models [Text] / J-H. Chang, N. S. Kim, S. K. Mitra // IEEE Trans. Signal Process. – 2006. – V. 54, № 6. – P. 1965–1976.
21. Харламов А. А. Анализ текстов: лингвистика, семантика, прагматика в рамках когнитивного подхода [Текст] / А. А. Харламов, Т. В. Ермоленко // Проблемы искусственного интеллекта. – 2015. – № 0 (1). – С. 106–115.

References

1. Jurafsky D., Martin J. H. *Speech and Language Processing*. 2nd Ed., NJ, Prentice Hall, 2009. 1024 p.
2. Flach P. *Machine learning. The art and science of algorithms that make sense of data.* Cambridge University Press, 2012. 410 p.
3. Rabiner L. R., Sambur M. R. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*,1975, vol. 54, no. 2, pp. 297–315
4. Van Gerven S., Xie F. A comparative study of speech detection methods. *Proceedings of European Conference on Speech, Communication and Technology*, Rhodos, 1997. Available at: http://www.mirlab.org/conference_papers/International_Conference/Eurospeech%201997/pdf/tab/a0199.pdf.
5. Marzinzik M., Kollmeier B. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech Audio Processing*, 2002, vol. 10, no. 2, pp. 109–118
6. Ramírez J., Segura J.C., Benítez C., De La Torre A., Rubio A. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 2004, vol. 42, pp. 271–287
7. Pencak J., Nelson D. The NP speech activity detection algorithm. *Proceedings of ICASSP*, Detroit, 1995. Available at: https://www.researchgate.net/profile/Douglas_Nelson8/publication/618329_The_NP_speech_activity_detection_algorithm/links/540857d90cf2c48563bb1228.pdf
8. Graf S., Herbig T., Buck M., Schmidt G. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015, vol. 2015, no.1, pp. 1-15
9. Nelson D. J., Pencak J. Pitch-based methods for speech detection and automatic frequency recovery. *Proceedings of SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation*, San-Diego, 1995. Available at: https://www.researchgate.net/ profile/Douglas_Nelson8/publication/260816047_Pitch-based_methods_for_speech_detection_and_automatic_frequency_recovery/links/541c15250cf2218008c4e563.pdf
10. Kristjansson T., Deligne S., Olsen P. Voicing features for robust speech detection. *Proceedings of INTERSPEECH*, Lisbon, 2005. Available at: http://papers.traustikristjansson.info/wp-content/uploads/2011/07/KristjanssonRobustVoicingEurospeech2005.pdf
11. Shahnaz C., Zhu W-P., Ahmad M O. A multifeature voiced/unvoiced decision algorithm for noisy speech. *Proceedings of ISCAS*, Kos, 2006, pp. 2528 -2531
12. Sadjadi S.O., Hansen J.H.L. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Processing*, Lett., 2013, vol. 20, no.3, pp. 197–200
13. Haigh J.A., Mason J.S. A voice activity detector based on cepstral analysis. *Proceedings of EUROSPEECH*, Berlin, 1993. Available at: https://www.semanticscholar.org/paper/A-voice-activity-detector-based-on-cepstral-HaighMason/0fc5b0a4d38a6ae1b5ce9bb347b82e3ef3505859/pdf
14. Kinnunen T., Rajan P. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. *Proceedings of ICASSP, Vancouver: IEEE*, 2013, pp. 7229–7233
15. Rabiner L. R. Sambur M. R. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Transactions on Speech Audio Processing*,1977, vol. 25, no.4, pp. 338–343
16. Ghosh P.K., Tsiartas A., Narayanan S. Robust voice activity detection using long-term signal variability. *IEEE Transactions on Speech Audio Processing*, 2011, vol. 19, no.3, pp. 600–613
17. Mesgarani N., Slaney M., Shamma S.A. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Speech Audio Processing*, 2006, vol. 14, no.3, pp. 920–930
18. Ezzat T., Bouvrie J., Poggio T. Spectro-temporal analysis of speech using 2-D Gabor filters. *Proceedings of INTERSPEECH*, Antwerp:ISCA, 2007. Available at: http://cbcl.mit.edu/projects/cbcl/publications/ps/ezzat-spetro-analysis-07.pdf
19. Sohn J., Kim N.S., Sung W. A statistical model-based voice activity detection. *IEEE Signal Processing*, Lett, 1999, vol. 6, no.1, pp. 1–3.
20. Chang J.-H., Kim N. S., Mitra S. K. Voice activity detection based on multiple statistical models. *IEEE Transactions on Signal Processing*, 2006, vol. 54, no.6, pp. 1965–1976
21. Kharlamov A. A., Ermolenko T. V. Analiz tekstov: lingvistika, semantika, pragmatika v ramkakh kognitivnogo podkhoda [Text analysis: linguistics, semantics, pragmatics within the cognitive approach]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2015, no. 0(1), pp. 106–115.

RESUME

T. V. Ermolenko, A. P. Tihonchuk

Determination of voice activity in speech

Background: For more than half a century people have been working at the ability to input data not only by hands, but also with the help of voice. The solution of the "voice recognition" problem can increase the efficiency of human-operator activity with machines. One of the most difficult points of searching for an answer is the detection of voice in the noise surrounding a microphone. This paper is aimed at disclosing one of the most promising ways to solve this problem.

Materials and methods: A number of voice activity detection (VAD) algorithms are applied, but for the most part they ignore the possibility of noise alternations. This makes previously prepared noise samples useless. To overcome this circumstance we need a distinctive feature that is peculiar only to speech. This should simplify the task and increase the efficiency of the voice recognition. We choose the entropy as a measure of disorder and therefore showing a very large robustness. For greater clarity the algorithm is tested on several speakers with different noise levels and voice characteristics.

Results: a set of data as a result of testing, and numerical performance indicators of the algorithm. The proposed algorithm also allows adjusting to the environment and optimizing the user's efforts. The analysis of the entropy application as a distinctive feature corroborates the above mentioned assumptions.

Conclusion: The applied algorithm shows fewer errors and higher speed with a low noise level, as well as acceptable functionality for a wide range of interference. The developed VAD-algorithm is expected to be used in the ASR-systems.

Статья поступила в редакцию 21.08.2017.