УДК 004.934.2

Г.А. ДОБРОВОЛЬСКИЙ, О.А. ТОДОРИКО
Запорожский национальный университет

ИСПОЛЬЗОВАНИЕ ЭНТРОПИИ ШЕННОНА ДЛЯ ДЕТЕКЦИИ ГОЛОСОВОЙ АКТИВНОСТИ В ЗАШУМЛЁННЫХ ЗВУКОЗАПИСЯХ

В работе рассматривается задача поиска человеческой речи в зашумленном звуковом сигнале. Проверяется гипотеза о том, что наличие речи увеличивает количество информации в соответствующих местах звукозаписи. Для проверки данной гипотезы был предложен способ использования энтропии Шеннона для формирования обучающей выборки простого классификатора. Проведенное исследование показало, что представленный способ позволяет обнаружить в звукозаписи человеческую речь даже если соотношение сигнал/шум не превышает 0 dB. Данный способ определения голоса предполагается использовать в системах обработки речи и автоматической проверки произношения при обучении иностранному языку.

Ключевые слова: обработка сигнала, обработка звука, информационная энтропия Шеннона, помехи, детектор речевой активности.

Г.А. ДОБРОВОЛЬСЬКИЙ, О.О. ТОДОРІКО
Запорізький національний університет

ЗАСТОСУВАННЯ ЕНТРОПІЇ ШЕННОНА ДЛЯ ВИЯВЛЕННЯ ГОЛОСОВОЇ АКТИВОСТІ В ЗАШУМЛЕНИХ ЗВУКОЗАПИСАХ

У роботі розглядається задача пошуку людської мови в зашумленому звуковому сигналі. Перевіряється гіпотеза про збільшення кількості інформації при наявності мови у відповідних місцях звукозапису. Для перевірки гіпотези було запропоновано спосіб використання ентропії Шеннона для формування навчальної вибірки простого класифікатора. Проведене дослідження показало, що представлений спосіб дозволяє виявити в звукозаписі людську мову навіть якщо співвідношення сигнал/шум не перевищує 0 dB. Даний спосіб визначення голосу передбачається використовувати в системах автоматичної обробки мови і автоматичної перевірки вимови під час навчання іноземній мові.

Ключові слова: обробка сигналів, обробка звуку, інформаційна ентропія Шеннона, шум, виявлення голосової активності

G.A. DOBROVOLSKY, O.A. TODORIKO
Zaporizhzhya national university

APPLICATION OF SHANNON ENTROPY FOR VOICE ACTIVITY DETECTION IN NOISY SOUND RECORDINGS

In this paper, the robust voice activity detection in noisy sound records is investigated. The hypotheses is tested that the value of Shannon information entropy rises if voice is added to noise. To verify the hypotheses, a method of training set preparation based on Shannon information entropy is proposed and implemented. The preliminary study shows that the proposed method allows voice detection even if signal to noise ratio equals to 0 dB or less. The proposed voice detection method is designed as part of speech processing software and automatic pronunciation scoring in computer assisted language learning.

Keywords: signal processing, sound processing, Shannon information entropy, noise, voice activity detection

Постановка проблемы

Серьезной помехой для большинство систем обработки речи, является шум. Автоматические телефонные сервисы, системы голосового управления, разнообразные модули автоматического улучшения качества звука и т.д. должны взаимодействовать с мобильными устройствами, которые зачастую используют микрофоны разного качества в разном звуковом окружении: дом, улица, работа, транспорт и т.д. Существующие способы выделения голоса и подавления помех часто требуют для обучения образцы "чистого шума" - фрагменты звука, не содержащие голоса. Но в условиях непредсказуемых помех признаки шума и голоса заранее неизвестны, что затрудняет дальнейшую обработку речи. Таким образом актуальной является задача классификации частей звукозаписи по признаку наличия/отсутствия речи в присутствии произвольных помех. Если учесть упомянутые выше условия, то задача классификации фрагментов звука по признаку наличия речи оказывается не столь тривиальной, как может показаться на первый взгляд, так как большинство детекторов голосовой активности срабатывают плохо, если уровень шума увеличивается.

Анализ последних исследований и публикаций

Типичная схема детектора голоса (рис.1) включает в себя модули извлечения признаков, классификации и сглаживания.

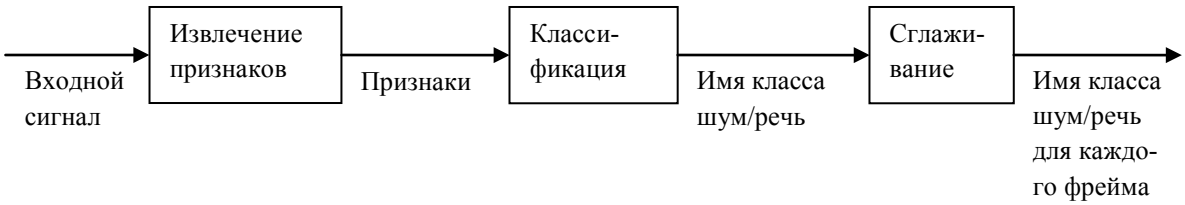

Рисунок 1. Схема детектора голоса

Целью модуля извлечения признаков является получение из входного сигнала числовых характеристик, упрощающих его классификацию. Входной сигнал является представлением звука в виде последовательности чисел (образцов), каждое из которых обозначает амплитуду $x(t_k)$ сигнала в часто расположенные моменты времени t_k. Интервал времени t_{k+1} - t_k определяется частотой дискретизации, которая обычно указывается в свойствах звукозаписи. Для вычисления признаков используются образцы, попадающие в заданный промежуток времени. Полученный набор образцов называется фреймом и состоит из K чисел

$$X(t_k) = < x_k, x_{k+1}, \dots , x_{k+K}> \quad (1)$$

Промежуток времени должен быть малым, чтобы сигнал в его пределах оставался практически постоянным, и одновременно содержать достаточное количество образцов, например, типичным значением для систем распознавания речи являются фреймы длиной 20-25 мс, расположенные с шагом 10-12,5 мс [1]. Модуль извлечения признаков создаёт на основе входного сигнала последовательность фреймов и вычисляет для каждого фрейма некоторый набор признаков.

Модуль классификации принимает на вход вычисленные признаки и возвращает на выходе идентификатор одного из существующих классов. В зависимости от природы признаков, для реализации модуля классификации обычно применяется один из методов машинного обучения [2], например, метод опорных векторов, нейронные сети различных архитектур, байесовские сети и т.д.

Модуль сглаживания позволяет уточнить результаты определения голоса на основании типичной длительности фонем. Например, единственный фрейм длины 25 мс, находящийся в окружении фонового шума, не может быть голосом, потому что в человеческой речи нет таких коротких фонем. Точно так же речевой аппарат человека не приспособлен к миллисекундным паузам в середине слова.

В настоящее время основным направлением совершенствования детекторов речи является выбор наиболее информативных признаков, позволяющих надёжно отличить произнесённые человеком слова от постороннего шума. Исторически первыми были попытки использовать энергию сигнала [3]. Этот простой признак хорошо работает для умеренного шума (сигнал/шум=30 дБ). Он основан на том факте, что в шумном окружении человек обычно говорит громче. Нормализация энергии увеличивает точность алгоритма. Также его можно адаптировать к медленному изменению шума. Недостатком метода является необходимость предварительно задать пороговое значение энергии. Усовершенствованными разновидностями детекторов голоса 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]. Сравнение выявило [8], что выделение фрейма с шумом и его дальнейшее сопоставление с другими фреймами [7], показывает наилучшие результаты.

Другая группа методов опирается на основную тональность звука. В соответствии с моделью произношения звуков человеком [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], мел-частотные 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. Недостатком подхода является необходимость обучения классификатора, что означает возрастание вероятности ошибки в условиях непредсказуемого шума.

Использование степени стационарности основано на наблюдении, что обычно шум изменяется намного медленнее, чем речь. 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. Недостатком такого подхода является возрастание ошибки в условиях нестационарного шума.

Ритм произнесения фонем. Ещё одна группа признаков опирается на ритмичность чередования согласных и гласных звуков в человеческой речи, что приводит к достаточно четким максимумам спектра в районе частоты 4 Гц. Признаки, использующие данное свойство, устойчивы к помехам. Однако для их вычисления необходимо рассматривать промежутки времени около 1 секунды. Например, спектрально-темпоральная модуляция (СТМ) [17] рассматривает модуляцию с изменением одновременно времени и частоты, стремясь смоделировать восприятие звука человеком, а также учитывает тональную и форматную структуру речи [18]. На сегодняшний день это один из самых надежных способов определения речи в звуке. К его 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.

Типичный статистический подход [19, 20] заключается в предположении, что звук голоса 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, также исследовался вопрос об их комбинации [21].

Приведённый выше анализ публикаций показывает, что возможность использования информационной энтропии Шеннона для детекции речи до сих пор не освещалась должным образом.

Цель работы

Целью данной работы является проверка возможности использования информационной энтропии Шеннона для определения голоса в звуковом сигнале.

Изложение основного материала исследования

Если считать речь носителем информации, то естественной будет гипотеза о том, что добавление речевого сигнала к фоновому шуму увеличит количество информации в соответствующих местах звукового сигнала. Для проверки данной гипотезы необходимо разработать способ проверки наличия голоса в звукозаписи и создать его программную реализацию. Общая структура реализации воспроизводит традиционное строение детектора голосовой активности, изображенное на рис.1.

Голосовая активность определяется по следующему сценарию: сначала проводится подготовка звукозаписи, после чего вычисляются признаки, выполняется поиск образцов шума, обучение классификатора, классификация фреймов по признаку наличия или отсутствия голоса, сглаживание.

График входного зашумленного сигнала для отношения сигнал/шум =0 dB показан на рис. 2.

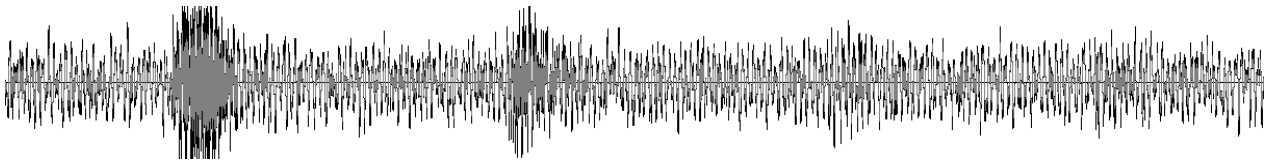

Рисунок 2. График входного зашумленного сигнала, содержащего текст “а б в г”

Подготовка звукозаписи выполняется традиционным для систем распознавания речи способом. Сначала с помощью усреднения удаляется низкочастотная составляющая сигнала, которая играет незначительную роль в распознавании речи:

$$x_k = \alpha\, x_k + (1 - \alpha)\; x_{r-1}\,,\ |\alpha|<1 \qquad (2)$$

где α - параметр, показывающий степень сглаживания.

После сглаживания производится линейное отображение амплитуды на отрезок [-1, 1] и разделение сигнала на фреймы. Для каждого фрейма вычисляются признаки: энергия, энтропия, MFCC [22].

Энергия вычисляется как смещённая оценка дисперсии входного сигнала:

$$E = \frac{1}{N}\sum_{k=0}^{N-1}\left(x_k - \bar{x}\right)^2 \; t=1..T, \tag{3}$$

где $\bar{x}$ - среднее значение сигнала, N - количество образцов в фрейме.

Для звукозаписи, изображенной на рис.2, количество энергии в каждом фрейме представлено на рис.3. Первый пик графика соответствует гласному звуку «а», второй – звонкому согласному «б», а пики для «в» и «г» практически незаметны на фоне шума.

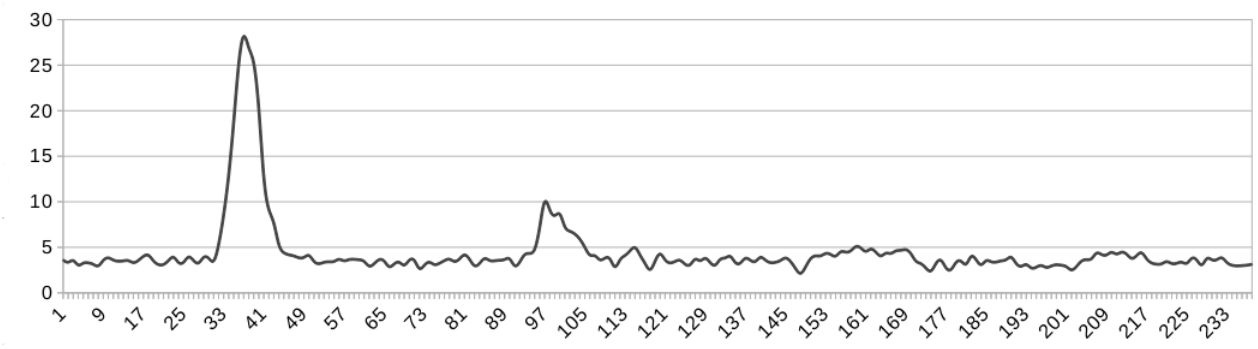

Рисунок 3. Количество энергии в каждом фрейме, горизонтальная ось соответствует номеру фрейма, вертикальная – количеству энергии

Для вычисления энтропии определяется размах амплитуды сигнала $[a_{min}, a_{max}]$, полученный отрезок делится на L частей $[a_0, a_1]$, $[a_1, a_2]$, … , $[a_{L-1}, a_L]$, где $a_0=a_{min}$ и $a_L=a_{max}$, и для каждого фрейма подсчитывается количество амплитуд, попавших в каждую из частей - составляется гистограмма частот. После этого по формуле Шеннона вычисляется информационная энтропия:

$$I = -\sum_{i=1}^{L} f_i \ln(f_i), \tag{4}$$

где f_i - доля амплитуд сигнала, попавших в промежуток $[a_{i-1}, a_i]$.

Для звукозаписи, изображенной на рис.2, значение энтропии каждого фрейма показано на рис.4. Горизонтальными полосками выделены участки с голосом, из рисунка видно, что перед началом и после окончания звука наблюдаются скачки значения энтропии.

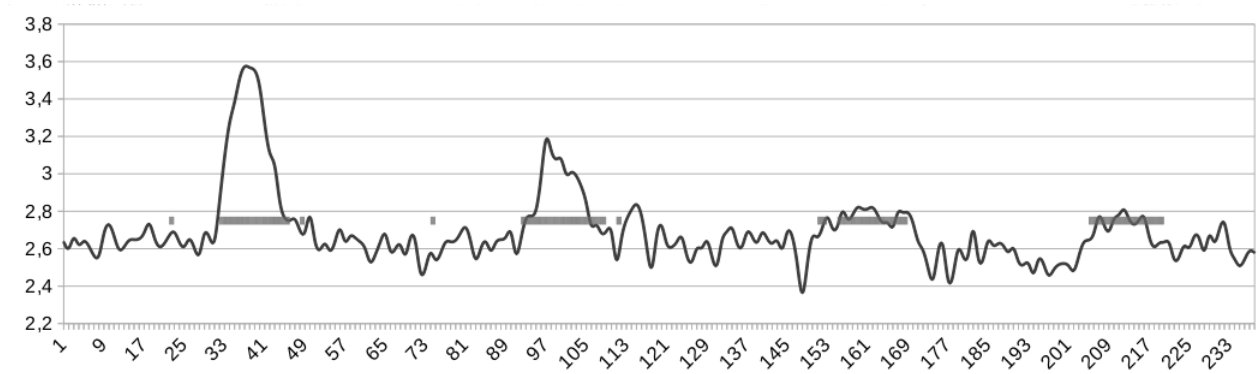

Рисунок 4. Значение энтропии для каждого фрейма, на горизонтальной оси отмечены номера фреймов, на вертикальной - значения информационной энтропии, горизонтальные полосы обозначают фреймы, в которых обнаружен голос.

Использование мел-частотных кепстральных коэффициентов (Mel Frequency Cepstral Coefficients, MFCC) считается стандартным методом извлечения признаков в системах распознавания речи [22]. Признаки MFCC вычисляются с помощью набора частотных фильтров, которые учитывают, что человеческий слух имеет разную чувствительность в разных частях звукового спектра – почти линейную для частот ниже 1 кГц и логарифмическую для более высоких частот.

На первом этапе находится логарифм энергии сигнала, который получается при наложении каждого фильтра

$$S(t,m) = \ln\left(\sum_{n=0}^{N-1}\left|X(t,n)\right|^2 H(m,n)\right), \mathrm{t} = 1..\mathrm{T}, \mathrm{m} = 0..\mathrm{M}-1 \tag{5}$$

где $X(t,n)$ – n-я компонента фурье-образа сигнала в фрейме t, $H(m,n)$ – n-я компонента m-го частотного mel-фильтра, N – размер окна и M – заданное заранее количество фильтров. Обычно в системах распознавания речи используют $M=20$, но $M=12$ тоже считается достаточным числом.

Второй этап сводится к дискретному косинусному преобразованию полученных значений $S(t,m)$:

$$c(t,m)=\sum_{m_1=0}^{M-1} S(t,m_1)\cos\left(\frac{m(m_1-0{,}5)\pi}{M}\right), t=1..T,\ m=0..M-1 \qquad (6)$$

Следующим шагом, после вычисления всех признаков, является поиск граничного значения энтропии, надёжно отделяющего фреймы, содержащие шум. Для этого гистограмма распределения значений энтропии всех фреймов приближается с помощью линейной комбинации двух нормальных распределений. Таким образом, учитываются свойства звукозаписи на больших отрезках времени. Использование именно двух нормальных распределений обусловлено предположением о скачкообразном изменении энтропии в моменты начала и окончания речи. Таким образом, гистограмма распределения энтропии должна иметь не менее двух максимумов. Это наглядно видно на рис. 5, где изображена гистограмма распределения значений энтропии для звукозаписи, изображенной на рис.2.

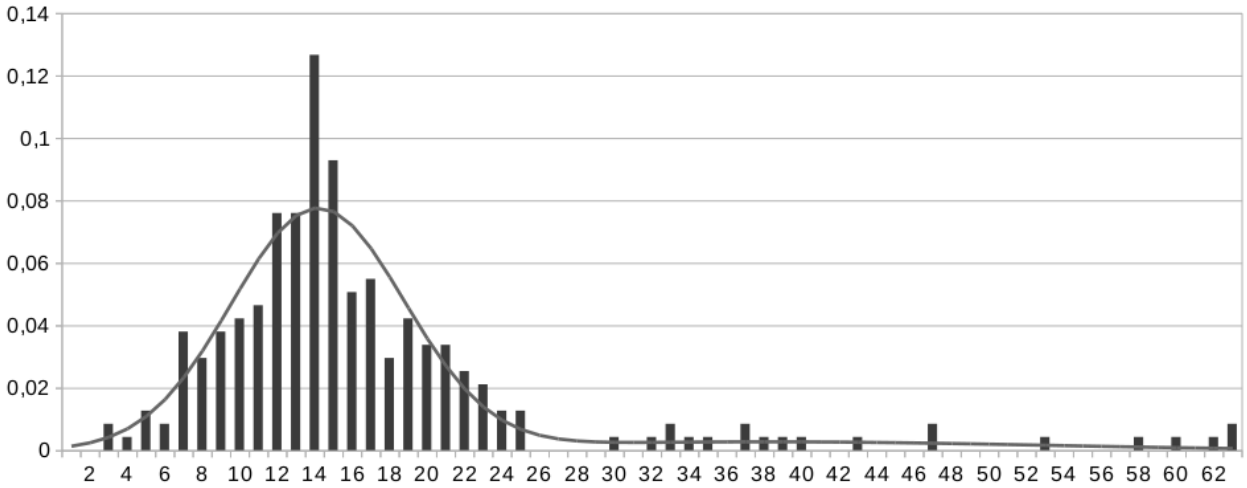

Рисунок 5. Гистограмма распределения энтропии и её аппроксимация с помощью суммы двух нормальных распределений, горизонтальная ось показывает номер интервала значений энтропии, вертикальная - относительное количество значений, попавших в каждый интервал.

С помощью EM-алгоритма [2] найденная гистограмма представляется как сумма двух нормальных распределений и находится минимальный экстремум <S> приближенной функции распределения. Фреймы, имеющие энтропию Шеннона, не превышающую <S>, считаются шумом и служат обучающей выборкой для классификатора.

При построении классификатора используется гипотеза, о том, что признаки MFCC подчиняются нормальному распределению. В таком случае обучение классификатора сводится к оценке значений математического ожидания и дисперсии для каждого признака MFCC. При классификации считается, что фрейм содержит голос, если среднее расстояние Махаланобиса от его набора признаков MFCC до найденного при обучении набора математических ожиданий превышает заданный априори предел.

Горизонтальные линии на рис. 4 показывают найденные фрагменты речи для неблагоприятного случая, когда сигнал по интенсивности равен шуму.

Выводы

В работе предложен способ использования энтропии Шеннона для формирования обучающей выборки простого классификатора. Предварительное исследование показало, что представленный способ позволяет даже в случае сильной зашумленности обнаружить в звукозаписи человеческую речь. Данный способ определения голоса предполагается использовать в системах автоматической обработки речи и автоматической проверки произношения при обучении иностранному языку.

Список использованной литературы

1. Jurafsky D. Speech and Language Processing (2Nd Edition) / Jurafsky, Daniel and Martin, James H. – NJ.: Prentice Hall, 2009 – 1024 pages
2. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / Петер Флах – М.: ДМК Пресс, 2015 – 400 с
3. Rabiner L.R. An algorithm for determining the endpoints of isolated utterances / 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 / 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 / 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 /J Ramírez, JC 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 / 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 / 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 / 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 / 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 / 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 / J.A.Haigh, J.S. Mason // Proc. of EUROSPEECH. – Berlin, 1993. – Режим доступа :
14. https://www.semanticscholar.org/paper/A-voice-activity-detector-based-on-cepstral-Haigh-Mason/0fc5b0a4d38a6ae1b5ce9bb347b82e3ef3505859/pdf
15. Kinnunen T. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data / T. Kinnunen, P. Rajan // Proc. of ICASSP. – Vancouver: IEEE, 2013 – P. 7229–7233
16. Rabiner L.R. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem / L.R. Rabiner, M.R. Sambur // IEEE Trans. Acoust. Speech Signal Process. – 1977. – V. 25, №4– P. 338–343
17. Ghosh P.K. Robust voice activity detection using long-term signal variability / P.K. Ghosh, A. Tsiartas, S. Narayanan // IEEE Trans. Audio, Speech, Lang. Process. – 2011. – V. 19, №3– P. 600–613
18. Mesgarani N. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations / N. Mesgarani, M. Slaney, S.A. Shamma // IEEE Trans. Audio, Speech Lang. Process. – 2006. – V. 14, №3– P. 920–930
19. Ezzat T. Spectro-temporal analysis of speech using 2-D Gabor filters / T. Ezzat, J. Bouvrie, T. Poggio // Proc. of INTERSPEECH. – Antwerp:ISCA, – 2007. – Режим доступа :
20. http://cbcl.mit.edu/projects/cbcl/publications/ps/ezzat-spetro-analysis-07.pdf
21. Sohn J. A statistical model-based voice activity detection / J. Sohn, N.S. Kim, W. Sung // IEEE Signal Process. Lett. – 1999. – V. 6, №1– P. 1–3.
22. Chang J.-H. Voice activity detection based on multiple statistical models. / .J-H Chang, NS Kim, SK Mitra // IEEE Trans. Signal Process. – 2006. – V. 54, №6– P. 1965–1976
23. Van Segbroeck M. A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice / M. Van Segbroeck, A. Tsiartas, S.S. Narayanan // Proc. of INTERSPEECH. – Lyon:ISCA, – 2013. – Режим доступа : https://www.semanticscholar.org/paper/A-robust-frontend-for-VAD-exploiting-contextual- Segbroeck-Tsiartas/d974bef7949dd95a848ba092e597ea1693c2e68d/pdf
24. Motlíček P. Feature Extraction in Speech Coding and Recognition, Report, Portland, to -research, data, and theory / Belmont, CA: Thomson/Wadsworth, 2003, – P. 1-50.