УДК 004.934.2 **DOI: 10.18413/2518-1092-2021-6-1-0-2**

Дьяченко А.В.
Подольский Д.А.

РЕАЛИЗАЦИЯ АЛГОРИТМА ДЕТЕКТИРОВАНИЯ АКТИВНОСТИ РЕЧИ ПРИ ПРОВЕДЕНИИ ПАРАЛИНГВИСТИЧЕСКОГО АНАЛИЗА

Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Кронверкский пр., д. 49, г. Санкт-Петербург, 197101, Россия

e-mail: ayrimur@mail.ru, podolsky.dmitry94@gmail.com

Аннотация
В настоящее время широко распространены алгоритмы детектирования активности речи. Такие алгоритмы нашли применение в различных задачах: при передаче речевого потока человека, в хранении информации (для компрессии аудиозаписей), для распознавания состояния человека при паралингвистическом анализе и т.д. Задача данной работы заключается в разработке и реализации алгоритма детектирования активности речи человека в программной среде Csound. На сегодняшний день уже существует ряд методов для распознавания речевой активности человека, такие как алгоритм определения скорости, метод адаптивного многоскоростного детектирования речи, метод, основанный на анализе спектральной формы и энергии поддиапазонов и т.д. [13, 16, 17], однако, на данный момент, указанные алгоритмы не были реализованы в среде Csound. В данной статье категоризированы признаки речи, описан реализованный алгоритм детектирования активности речи, а именно определения пауз паралингвистического анализа речевого аудио с использованием преобразования Гильберта, что уменьшает сложность алгоритма, сохраняя его точность. Целью данной работы является модификация и реализация алгоритма обнаружения речевой активности в помещении на основе речевого потока в среде Csound для проведения паралингвистического анализа речевой активности человека.
Ключевые слова: детектор активности речи; Csound; паралингвистический анализ; активность речи.
Для цитирования: Дьяченко А.В., Подольский Д.А. Реализация алгоритма детектирования активности речи при проведении паралингвистического анализа // Научный результат. Информационные технологии. – Т.6, №1, 2021. – С. 13-19. DOI: 10.18413/2518-1092-2021-6-1-0-2

Diachenko A.V.
Podolsky D.A.

IMPLEMENTATION OF THE SPEECH ACTIVITY DETECTING ALGORITHM AT CONDUCTING PARALINGUISTIC ANALYSIS

Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, 49 Kronverkskiy prospekt, St. Petersburg, 197101, Russia

e-mail: ayrimur@mail.ru, podolsky.dmitry94@gmail.com

Abstract
Algorithms for the speech activity detecting are now widely used. Such algorithms are used in various tasks: transmitting a human speech stream, storing information for compressing audio recordings, for recognizing a person's state in the paralinguistic analysis, etc. The goal of this work is to develop and implement an algorithm for detecting human speech activity using the Csound software environment. Recently, there are already a number of methods for human speech activity recognition, such as the speed determination algorithm, the adaptive multi rate speech detection method, the method based on the analysis of the spectral shape and energy of subbands, etc. [13, 16, 17], however, at the moment, these algorithms haven't been implemented in the Csound environment. This article categorizes speech features, describes an implemented algorithm for detecting speech activity, namely, determining pauses in paralinguistic analysis of speech audio using the Hilbert transform, which reduces the complexity of the algorithm, while

maintaining its accuracy. The aim of this work is to modify and implement an algorithm for detecting speech activity in a room based on the speech flow in the Csound environment for conducting paralinguistic analysis of human speech activity.
Keywords: voice activity detection; Csound; paralinguistic analysis; speech activity.
For citation: Diachenko A.V., Podolsky D.A. Implementation of the speech activity detecting algorithm at conducting paralinguistic analysis // Research result. Information technologies. – T.6, №1, 2021. – P. 13-19. DOI: 10.18413/2518-1092-2021-6-1-0-2

ВВЕДЕНИЕ

Современные исследования по обработке речи направлены на решение вопросов её интерпретации и распознавании средствами компьютерного анализа и машинного обучения. Алгоритмы, разрабатываемые в данной предметной области, могут быть направлены на определение и распознавание отдельных слов и слитной речи (высказываний и фраз), определение эмоционального состояния диктора, диагностику патологий голосового аппарата, автоматический перевод, идентификацию личности по голосу, обучение иноязычной речи и др. [5].

Существующие алгоритмы распознавания эмоций по речевому сигналу не способны определять текущее психофизическое состояние диктора с высокой точностью. Это, в первую очередь, обусловлено сложностью определения эмоций как таковых, а также необходимостью конечного выбора частотно-временных характеристик аудиопотока, в качестве анализируемых параметров алгоритма [7].

Реализуемый в данной работе алгоритм выделяет из речевого потока векторы с невербальной информацией потока, а именно – вектор пауз и активности речи диктора, которые, в свою очередь, планируется передавать в системы определения эмоционального состояния, в том числе и на основе искусственного интеллекта.

ОСНОВНАЯ ЧАСТЬ

В общем случае речевой сигнал можно рассматривать как набор вербальных и невербальных составляющих речи. Для анализа вербальной составляющей потока применяют преимущественно лингвистические методы, а для невербальной – паралингвистические и экстралингвистические [1]. Лингвистические методы обработки речевого сигнала основываются на определении семантических единиц речи, а именно: звуков, букв, слов, которые произносит человек. Паралингвистический и экстралингвистический анализ позволяет судить обо всех звуковых составляющих речи – интонация, темп и мелодика речи, паузы во время коммуникации и др.

Невербальную составляющую речи изучает паралингвистика. С помощью паралингвистического подхода можно выявить эмоции, интонации, психофизиологические состояния, особенности произношения и прочие параметры голоса диктора [6]. Согласно исследованиям [2], информация, передаваемая людьми вербально, составляет лишь около 7% общего объёма передаваемой информации – остальные почти 93% полезной информации передаётся невербальным способом – с помощью мимики, поз, жестов, движений и т.д. При этом не менее трети всей информации приходится на паралингвистическую составляющую речи. Паралингвистическую информацию также обычно отличают от экстралингвистической, с которой ассоциируют не связанные с речью акустические явления, например, атипичные особенности речи: речевые паузы, вздохи, заикание, покашливание и пр. (рис 1). Таким образом, с помощью паралингвистического анализа речевого информационного канала, можно определить состояние человека, его настроение и даже патологии речевого тракта или участков мозга, отвечающих за речевой информационный обмен [3].

В работе [10] подробно описывается, что частотный диапазон сигнала 200-300 Гц несёт в себе информацию преимущественно о вербальной составляющей речи, а диапазон частот свыше 300 Гц – о невербальной составляющей сигнала, т.е. Сигнал в диапазоне ниже 300 Гц не будет нести смысловой информации, однако будет содержать данные об эмоциональном состоянии говорящего.

В рамках данной статьи описан реализованный алгоритм детектирования активности речи, а именно определения пауз паралингвистического анализа речевого аудио.

Речь человека может быть представлена как блоки информации, перемежающиеся блоками пауз – в ней присутствуют характерные паузы между семантическими единицами, в отличии от других сигналов [15].

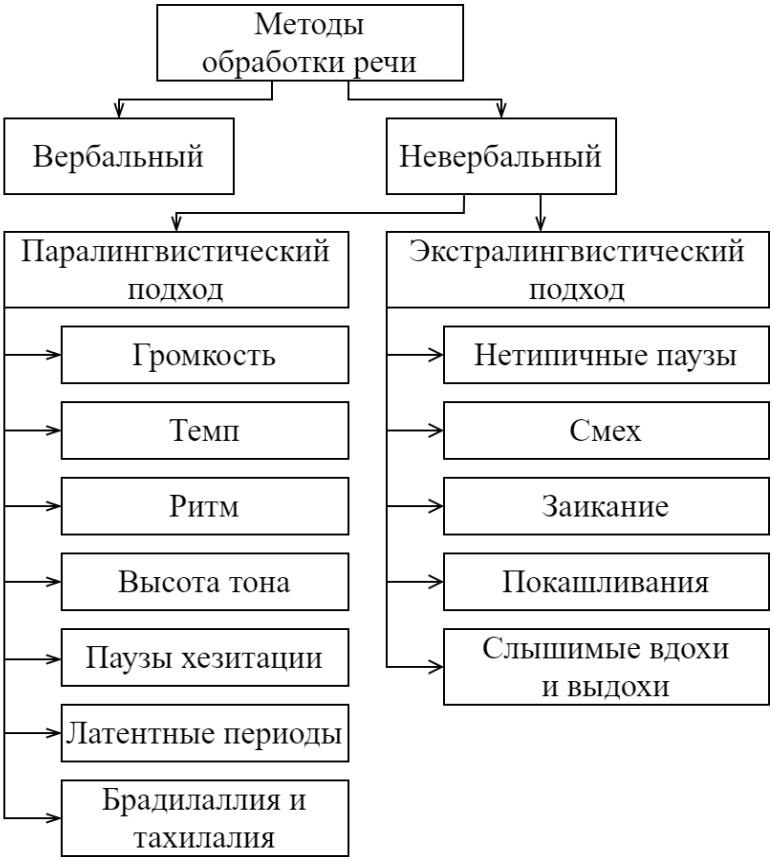

Рис. 1. Категоризация признаков речи
Fig. 1. Categorization of speech features

Определение активности речи уже широко применяется для сжатия сигнала, сокращения объема передаваемой информации в телекоммуникационных системах по каналам с ограниченными пропускными способностями. Но, кроме этого, активность речи – одна из характеристик, несущая информацию об эмоциональном состоянии говорящего человека. Активность определяется посредством обнаружения пауз в речи (рис. 2) – их длительность, частоту. Чем чаще паузы – тем выше активность речи, чем реже – тем ниже.

Так, определив уровень активности речи, можно сделать вывод о возрасте или о темпераменте человека, а, проследив за этой характеристикой в динамике, можно сказать об относительном состоянии человека в данной ситуации (например, взволнованность или возбужденность обычно сопровождаются повышением активности речи, а усталость, плохое самочувствие или неуверенность – снижением таковой) [3].

Для распознавания активности речи аудиопоток делят на участки активной речи и пауз. При этом участками активной речи называют те интервалы, где присутствует голосовая информация, а интервалы между этими участками называют участками пауз.

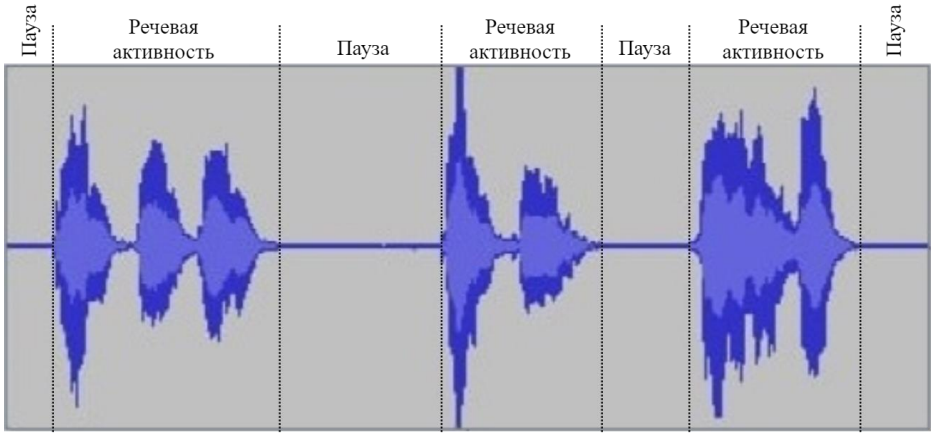

Рис. 2. Осциллограмма с выделенными участками активной речи и пауз
Fig. 2. Oscillogram with highlighted areas of active speech and pauses

Основой любого метода определения пауз служит детектор активности речи (voice activity detector – VAD) – это алгоритм, предназначенный для различения интервалов активной речи и пауз [11]. Чаще всего встречаются детекторы активности речи, алгоритмы которых работают на основе анализа энергии сигнала, обнаружения основного тона, спектрального и кепстрального анализа, измерений числа переходов через нуль, анализе энергии поддиапазонов, статистического моделирования, использования порядковых фильтров и т.д. [13, 16, 17].

Подробное описание взятого за основу для реализации алгоритма приведено в статье [8]. Суть описанного алгоритма сводится к локализации пауз методом цифровой фильтрации в двух спектральных диапазонах, соответствующих локализации максимумов энергии для вокализованных (диапазон частот 150-1000 Гц) и шумных (или невокализованных) звуков (диапазон частот 1500-3500 Гц) полосовыми фильтрами, “взвешивания” кратковременной энергии речевого сигнала в этих диапазонах с использованием прямоугольного окна. Структурная блок-схема реализованного алгоритма представлена на рисунке 3.

Рис. 3. Блок-схема алгоритма
Fig. 3. The block diagram of the algorithm

Для реализации выбранного алгоритма был использован специализированный язык программирования для синтеза звука и обработки звуковых сигналов Csound. Csound – высокоуровневый компьютерный язык программирования [14, 18], который является кросс-платформенной средой с открытым исходным кодом, поддерживающий все современные операционные системы. Кроме того, он может работать с платформами Raspberry Pi, BeagleBone и Arduino. Csound позволяет синтезировать звуки как в режиме реального времени, так и в автономном режиме, а также обрабатывать аудио в частотной и временной областях. Csound обладает большой библиотекой инструментов для синтеза и работы со звуком, в том числе фильтры и инструменты для спектральной обработки аудиосигналов.

Исходный входной речевой сигнал нормируется для выравнивания громких и тихих участков с целью исключения зависимости результатов от уровня входного сигнала. Нормализация сигнала возможно только для уже записанных (стационарных) или детерминированных (с точки зрения амплитудных характеристик) речевых сигналов. В противном случае операция нормирования сигнала может быть осуществлена на участках относительно локального максимум, при этом необходимо учитывать зависимость конечного результата работы алгоритма.

Из нормализованного сигнала полосовыми фильтрами выделяются два частотных диапазона. В Csound реализован полосовой фильтр второго порядка, который с помощью опкода *resonr* выделяет диапазоны 150-1000 Гц и 1500-3500 Гц.

Выделенные частотные полосы поступают в блок преобразования Гильберта. Последнее традиционно используется для выделения амплитудной огибающей сигнала на основе квадратурной пары сигналов (1):

$$\hat{S}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{S(\tau)}{t-\tau} d\tau, \quad (1)$$

где S(τ) – исходный сигнал.

Огибающей пары сигналов {s, ŝ} называют неотрицательную функцию времени (2):

$$S(t) = |(s^2 + \hat{s}^2)^{0,5}|, \quad (2)$$

Сигналы огибающих, полученных для каждой полосы, поступают на вход решающего блока, в котором устанавливаются пороговые значения (уровни) срабатывания для определения наличия паузы в исходном сигнале. Пороговые значения K1 и K2 в рамках описываемой реализации устанавливаются эмпирически, в процессе первоначальной калибровки работы алгоритма.

Реализованный алгоритм детектирования активности речи позволяет определять активные участки речи, а также считать количество пауз в заданном временном интервале. Использование блока преобразования Гильберта позволило избежать реализацию вычисления энергии сигнала в выделенных частотных полосах, а также, за счет преимуществ преобразования, работать со слабо детерминированными речевыми сигналами.

Csound поддерживает редактирование формата выходных векторов с блоков счетчика пауз и детектора тишины. Это позволяет интегрировать реализованный алгоритм в комплексы паралингвистического анализа, которые в том числе используют алгоритмы искусственного интеллекта для минимизации ошибки при калибровке параметров решающего блока.

ЗАКЛЮЧЕНИЕ

В результате выполнения работы, был реализован алгоритм для распознавания речевой активности с помощью анализа двух диапазонов частот и преобразования Гильберта в среде Csound. На выходе реализованного алгоритма имеется два вектора – с блоков счетчика пауз и детектора тишины. Это позволяет интегрировать реализованный алгоритм в комплексы паралингвистического анализа, которые в том числе используют алгоритмы искусственного интеллекта для минимизации ошибки при калибровке параметров решающего блока.

Список литературы

1. Айвазян О.О. Вербальная и невербальная коммуникация, как факторы развития речи // Конференция "Стратегические направления устойчивого развития социально-экономической политики южного региона". Майкоп, 2012.

2. Басов О.О., Карпов А.А., Саитов И.А. Методологические основы синтеза полимодальных инфокоммуникационных систем государственного управления. Орёл: Академия ФСО России, – 2015. – 271 с.

3. Василик. М.А. Пара- и экстралингвистические особенности невербальной коммуникации // Элитариум, 2018. URL: www.elitarium.ru/neverbalnoe-obshhenie-temp-rech-golos-informacija-kommunikacija-intonacija-vnimanie (дата обращения: 16.12.2020).

4. Величко А.Н., Будков В.Ю., Карпов А.А. Аналитический обзор компьютерных паралингвистических систем для автоматического распознавания лжи в речи человека // Информационно-управляющие системы. – 2017. – №5(90). – С. 30-41.

5. Вердербер, Р., К. Вердербер. Психология общения: Тайны эффектив. Взаимодействия. М.: Прайм-ЕВРОЗНАК: Олма-Пресс, 2003. – 320 с.

6. Карпов А.А., Кайа Х., Салах А.А. Актуальные задачи и достижения систем паралингвистического анализа речи // Научно-технический вестник информационных технологий, механики и оптики. – 2016. – Т. 16. – № 4. – С. 581–592.

7. Потапова Р.К., Бобров Н.В. Основные тренды в развитии междисциплинарной концепции “Анализ-синтез-анализ речи” // Математические методы в технике и технологиях. – 2019. – Т.7. – С. 124-129.

8. Практический алгоритм определения темпа речи для использования в контактцентрах / Никифоров С.Н., Никифоров Д.С., Виторский И.И., Танюкевич М.С. // Речевые технологии. – 2020. – № 1. – С. 6-12.

9. Симончик К.К., Галинина О.С., Капустин А.И. Алгоритм обнаружения речевой активности на основе статистик основного тона в задаче распознавания диктора // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика, телекоммуникации и управление. – 2010. – №4(103). – С. 23-31.

10. Чухрова М.Г. Взаимосвязь психоэмоционального состояния младших школьников и их голосоречевых характеристик // Наука и социум / Мат. Всерос. Науч.- практ. конф. с междунар. участием. 1 марта 2018 г. – Новосибирск: ЧУДПО 2018. – С. 99-104.

11. Шелухин О.И., Лукьянцев В.Г. Цифровая обработка и передача речи. М.: Радио и связь, 2000. – 456 с.

12. Волченков В.А., Витязев В.В. Методы и алгоритмы использования детектирования активности речи // Цифровая обработка сигналов. – 2013. – №1. – С. 54-60.

13. Adil Benyassine, H.Y. Eyal Shlomot, Dominique Massaloux Su. Silence compression scheme for use with g. 729 // Digital simultaneous voice and data applications IEEE Commun. Mag., 1997. – №35(9). – P. 64-73.

14. Boulanger R. The Csound Book: Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming. Cambridge: MIT Press, 2000. – 782 p.

15. Kondoz A.M. Digital Speech. Coding for Low Bit Rate Communication Systems. John Wiley & Sons, Ltd. 2004. – 442 p.

16. Prasad R. Comparison of Voice Activity Detection Algorithms for VoIP // In proc. 7th IEEE symp. on Computer science. 2005. – p. 567-576.

17. Sunil Kumar S.B., Sreenivasa Rao K. Voice/non-voice detection using phase of zero frequency filtered speech signal // Speech Communication. – 2016. – №81. – P. 90-103.

18. Vercoe B. Csound: A Manual for the Audio-Processing System. MIT Media Lab, 1995. – 341p.

References

1. Ayvazyan OO Verbal and non-verbal communication as factors of speech development // Conference "Strategic directions of sustainable development of socio-economic policy of the southern region". Maykop, 2012.

2. Basov OO, Karpov A.A., Saitov I.A. Methodological foundations for the synthesis of polymodal infocommunication systems of public administration. Oryol: Academy of the FSO of Russia, 2015. – 271 p.

3. Vasilik. M.A. Para- and extralinguistic features of non-verbal communication // Elitarium. 2018. URL: www.elitarium.ru/neverbalnoe-obshhenie-temp-rech-golos-informacija-kommunikacija-intonacija-vnimanie (date of access: 16.12.2020).

4. Velichko A. N., Budkov V. Yu., Karpov A. A. Analytical review of computer paralinguistic systems for automatic recognition of lies in human speech // Information and control systems. 2017. – No. 5(90). – P. 30-41.

5. Werderber, R., K. Werderber. Psychology of communication: Secrets are effective. Interactions. M.: Prime-EVROZNAK: Olma-Press, 2003. – 320 p.

6. Karpov A.A., Kaya H., Salah A.A. Actual problems and achievements of systems for paralinguistic speech analysis // Scientific and technical bulletin of information technologies, mechanics and optics. 2016. – Vol. 16. – No. 4. – PP. 581-592.

7. Potapova R.K., Bobrov N.V. Main trends in the development of the interdisciplinary concept “Analysis-synthesis-analysis of speech” // Mathematical methods in engineering and technology. 2019. – T.7. – P. 124-129.

8. Practical algorithm for determining the rate of speech for use in contact centers / Nikiforov S.N., Nikiforov D.S., Vitorsky I.I., Tanyukevich M.S. // Speech technologies. 2020. – No. 1. – P. 6-12.

9. Simonchik K.K., Galinina O.S., Kapustin A.I. Algorithm for detecting speech activity based on pitch statistics in the problem of speaker recognition // Scientific and technical bulletins of the St. Petersburg State Polytechnic University. Computer science, telecommunications and management. 2010. No. 4 (103). P. 23-31.

10. Chukhrova M.G. The relationship between the psychoemotional state of primary schoolchildren and their voice-speech characteristics // Science and society / Mat. Vseros. Scientific – practical. conf. with int. participation. March 1, 2018 – Novosibirsk: CHUDPO 2018. – P. 99-104

11. Shelukhin O.I., Lukyantsev V.G. Digital processing and transmission of speech. M.: Radio and communication, 2000, 456 p.

12. Volchenkov V.A., Vityazev V.V. Methods and algorithms for using speech activity detection // Digital signal processing. – 2013. – No. 1. – P. 54-60.

13. Adil Benyassine, H.Y. Eyal Shlomot, Dominique Massaloux Su. Silence compression scheme for use with g. 729 // Digital simultaneous voice and data applications IEEE Commun. Mag., 1997. – №35(9). – P. 64-73.

14. Boulanger R. The Csound Book: Perspectives in Software Synthesis, Sound Design, Signal Processing, and Programming. Cambridge: MIT Press, 2000. – 782 p.

15. Kondoz A.M. Digital Speech. Coding for Low Bit Rate Communication Systems. John Wiley & Sons, Ltd. 2004. – 442 p.

16. Prasad R. Comparison of Voice Activity Detection Algorithms for VoIP // In proc. 7th IEEE symp. on Computer science. 2005. – p. 567-576.

17. Sunil Kumar S.B., Sreenivasa Rao K.. Voice/non-voice detection using phase of zero frequency filtered speech signal // Speech Communication. – 2016. – No 81. – P. 90-103.

18. Vercoe B. Csound: A Manual for the Audio-Processing System. MIT Media Lab, 1995. – 341p.

Дьяченко Анна Витальевна, студент Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики
Подольский Дмитрий Анатольевич, инженер, Национальный центр когнитивных разработок Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики

Diachenko Anna Vitalievna, student, Saint Petersburg National Research University of Information Technologies, Mechanics and Optics
Podolsky Dmitry Anatolievich, engineer, ITMO University's National Center for Cognitive Technologies