# КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 001.891.573 : 004.934.2 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait/1995-5499/2022/4/180-188
Поступила в редакцию 14.07.2022
Подписана в печать 05.12.2022

# МЕТОД АНАЛИЗА РЕЧЕВОГО СИГНАЛА ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ АГРЕССИИ В РАЗГОВОРНОЙ РЕЧИ

**© 2022 А. Н. Величко**✉

*Санкт-Петербургский Федеральный исследовательский центр Российской академии наук*
*14 линия, 39, 199178 Санкт-Петербург, Российская Федерация*

**Аннотация.** В последние годы все более актуальной становится тема определения деструктивного поведения людей в сети Интернет для обеспечения их психологического комфорта. Деструктивное поведение является разрушительным поведением, а агрессия в европейской культуре представлена как мотивированное деструктивное поведение, которое может быть направлено как вовне, так и на себя, а также противоречит общепринятым социальным нормам. Данная работа рассматривает агрессию как паралингвистические явление, то есть, то, как агрессия проявляется в речи, а не то, что именно человек говорит. В статье представлены понятие и виды агрессии, приведен краткий анализ существующих работ. Представлена формальная постановка мультиклассовой задачи классификации и описание предложенного метода определения агрессии в речи. Были проведены представлены экспериментальные исследования методов классификации для автоматического определения агрессии, где лучшим оказался метод случайного леса, поскольку с его помощью удалось получить наилучшие и наиболее стабильные результаты. На основе полученных экспериментальных исследований был разработан предлагаемый метод определения агрессии в разговорной речи. Были использованы многомодальные корпуса Stress at Service Desk Dataset и Aggression in Trains, из которых были извлечены аудио дорожки для обучения и тестирования моделей с использованием 5-кратной перекрестной валидации. Предложенный метод представляет собой ансамбль из методов случайного леса, обученных на различных наборах акустических признаков с различными весами. Лучший результат, полученный с использованием предложенного метода равен 76,5 % по показателю невзвешенной средней полноты, и является одним из лучших среди аналогичных методов определения агрессии в разговорной речи.

**Ключевые слова:** компьютерная паралингвистика, деструктивное поведение, агрессия, автоматическое определение агрессии в разговорной речи, речевые технологии, мультиклассовая классификация, машинное обучение.

## ВВЕДЕНИЕ

В последние годы возрос уровень деструктивного поведения людей, что в большой степени проявляется при виртуальной коммуникации в сети Интернет. В связи с этим является актуальной тема выявления деструктивных (девиантных, агрессивных и враждебных) действий пользователей и обеспечение психологического комфорта пользователей.

Паралингвистика изучает то, как речь произносится, а не то, что именно произносится [1]. Для паралингвистики при выявлении невербальных характеристик в речи человека голосовые характеристики являются более важными, чем слова. В этом случае наи-

✉ Величко Алёна Николаевна
e-mail: alena.n.velichko@gmail.com

более распространенными акустическими признаками считаются основной тон, частота основного тона, паузы и т. п.

Под деструктивным поведением в психологии понимают разрушительное поведение человека, которое может быть направлено как вовне, так на самого человека, и может проявляться вербальным или практическим методом. Под девиантным поведением чаще всего подразумевается поведение личности, которое отклоняется от общепринятой нормы, распространенных, устоявшихся и общественных норм. То есть данная дефиниция базируется на понятии «норма».

Под термином агрессия в европейской культуре подразумевается деструктивное поведение, которое является мотивированным, а также противоречит нормам сосуществования людей. Такое поведение может быть направлено вовне, и иногда на себя. Согласно опроснику Басса — Дарки существует несколько видов агрессивных реакций [2]:

• Физическая агрессия — применение физической силы против собеседника.

• Косвенная агрессия — непрямым путем направленная на собеседника, или не направленная ни на кого.

• Раздражение — готовность к проявлению негативных чувств при малейшем возбуждении.

• Обида — ненависть или зависть к собеседнику по существующей или надуманной причине.

• Подозрительность — может находиться в интервале от недоверия и осторожности по отношению к окружающим до убежденности в том, что окружающие люди хотят нанести или наносят вред.

• Вербальная агрессия — вербальное проявление негативных чувств как через крик или визг, так и через словесные ответы.

• Чувство вины — выражение возможного убеждения субъекта в том, что он плохой человек и поступает плохо, субъект также ощущает угрызения совести.

Графически формы агрессивного поведения по классификации Басса представлены на рис. 1.

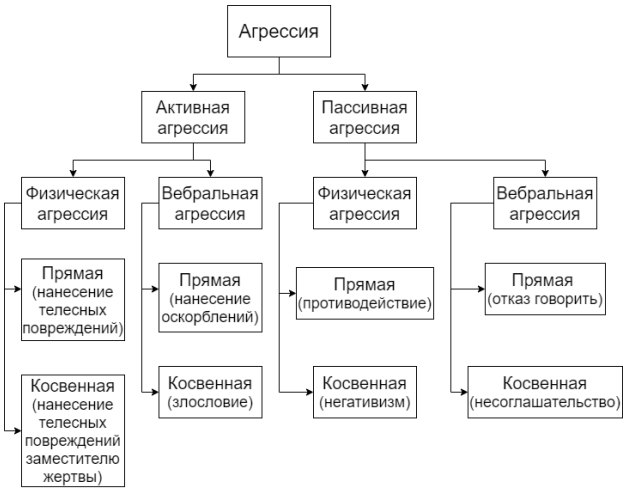


*Рис. 1. Классификации видов агрессии по А. Бассу (по классификации [2])*
*[Fig. 1. Aggression types classification according to A. Buss (based on the classification [2])]*

Аутоагрессия является причинением субъектом вреда себе (как физического, так и психологического) и относится к механизмам психологической защиты. Аутоагрессия может проявляться в самоунижении, самообвинении, нанесении себе телесных повреждений вплоть до самоубийства. К аутоагрессии также относится саморазрушительное поведение (алкоголизм, наркомания, выбор экстремальных видов спорта, опасных профессий, провоцирующее поведение). Аутоагрессия считается типичной для депрессивных личностей, также она может быть свойственна людям с мазохистическим характером [9].

В работе [13] было выявлено, что характеристиками агрессии являются такие акустические показатели как: высокая громкость речи и ее высокая вариативность, низкая высота основного тона и ее высокая вариативность, быстрая скорость речи, короткая длительность речи и короткая длительность пауз, а также малое количество пауз.

В 2021 году в рамках международной конференции INTERSPEECH на соревнованиях ComParE была представлена тема определения уровня агрессии. В качестве данных участникам был предложен набор данных, состоящий из двух речевых корпусов: Dataset of Aggression in Trains (TR) [5] и the Stress at Service Desk Dataset (SD) [7]. Всего для обучения, отладки и тестирования моделей

было предложено 911 аудио записей. Помимо набора данных были также предоставлены несколько наборов акустических признаков. С использованием набора признаков ComParE 2013 и метода опорных векторов организаторами конкурса был получен базовый результат 72,2 % по показателю невзвешенной средней полноты (unweighted average recall, UAR).

Победители данного соревнования [4] использовали X-вектора, вычисленные из спектрограмм и базовые признаки, а также векторы Фишера. Для классификации использовался метод опорных векторов, обучение каждого классификатора сначала проходило на отдельных наборах признаков, а затем предсказания объединялись. Авторам удалось добиться результатов на тестовом наборе данных в 61,5 % по показателю UAR при использовании комбинации признаков с X-векторами и 63,2 % при использовании комбинации признаков с векторами Фишера. На отладочном наборе авторы добились результата UAR = 77,8 %.

В работе [6] авторы предлагают распознавать агрессию при помощи определения наложения речи (перебивания) субъектов. Авторами была разработана система терапии с использованием виртуальной реальности, целью которой является помощь пациентам судебно-медицинских клиник для борьбы со склонностью к агрессии. Помимо акустических признаков был использован вектор признаков, состоящий из информации о наложениях речи, представленных тремя категориями: короткий ответ, преждевременный коммуникативный ход и состязательное перебивание оппонента. В качестве метода классификации был выбран метод случайного леса. Авторы отметили, что использование наложений речи позволило улучшить результат определения агрессии в речи на 3 % до 53,0 % по показателю невзвешенной точности (Unweighted Accuracy, UA).

Авторы работы [11] для определения агрессии в речи применяли признаки изменения давления воздуха в разных отделах голосового тракта. Они выявили, что одни и те же гласные, произнесенные с агрессией и без показывают различные значения давления воздуха в разных отделах голосового тракта. Для уменьшения размерности признакового пространства был применен метод главных компонент, а для классификации использовался метод скрытых марковских моделей. Сначала выносилось решение для каждой гласной в аудио записи, после чего итоговое предсказание производилось за счет голосования по большинству. Авторам удалось добиться результата распознавания по показателю точности (Accuracy, Acc) до 93,0 %.

В работе [15] использовали акустическую и лексическую информацию для определения агрессии. Авторы использовали несколько наборов данных для экспериментов: TR и SD, RAVDESS, CREMA-D, SAVEE и TESS. Сначала авторы применили метод обнаружения голосовой активности во входном акустическом сигнале (Voice Activity Detector, VAD) для удаления сегментов, не содержащих речь. Затем из аудио данных были извлечены MFCC признаки, которые впоследствии были поданы на вход предобученной нейронной сети ResNet-18. Для лексической составляющей был использован предобученный метод на основе энкодеров (Sentence-BERT, SBERT). На последнем этапе для классификации уровня агрессии был выбран метод опорных векторов. Лучшим результатом, полученным с использованием такого подхода, является результат 81,5 % по показателю UAR для TR и SD наборов данных при использовании переноса обучения с наборов данных RAVDESS, CREMA-D, SAVEE и TESS.

## 1. ПРЕДЛАГАЕМЫЙ ПОДХОД К ОПРЕДЕЛЕНИЮ АГРЕССИИ В РАЗГОВОРНОЙ РЕЧИ

Формальная постановка задачи может быть представлена следующим образом. Пусть $X$ — множество объектов. Признак — это, по сути, результат измерения какой-то характеристики объекта, его можно выразить как отображение этого объекта

$$f : X \to D_f,$$

где $D_f$ — множество допустимых значений признака. В нашем случае чаще всего это мно-

жество равно множеству действительных чисел (т. е. признаки являются количественными).

Если заданы признаки $f_1,\ldots,f_n$, то вектор $x=(f_1(x),\ \ldots,f_n(x))$ будет признаковым описанием объекта $x\in X$. В области машинного обучения допускается отожествление признакового пространства с самими объектами, т. е. множество $X$ вида

$$X=D_{f1}\times D_{f2}\times\ldots\times D_{fn},$$

где × обозначает декартово умножение, можно назвать признаковым пространством.

Таким образом, матрица объектов-признаков будет представлять совокупность признаковых описаний объектов обучающей выборки $X^l=(x_1,x_2,\ \ldots\ ,x_l)$ длины $l$, которая записана в виде матрицы размера $l\times n$ ($l$ строк, $n$ столбцов), столбцы которой соответствуют признакам $f_1,\ldots,f_n$, а строки — признаковым описаниям одного обучающего объекта.

Пусть имеется множество описаний объектов $X$ и множество номеров классов $Y$. Существует неизвестная целевая зависимость — отображение $y^*: X\to Y$, при этом ее значения известны только на объектах конечной обучающей выборки $X^m=\{(x_1,y_1),\ldots,(x_m,y_m)\}$. Тогда требуется найти алгоритм $a: X\to Y$, который сможет классифицировать объекты x множества $X$.

В нашем случае необходимо найти алгоритм

$$a_{agg}: X^m_{agg}=\{(x_1,y_1),\ldots,(x_m,y_m)\}\to Y_{agg},$$

где входные данные представлены матрицей объектов-признаков $X^l=(x_1,x_2,\ldots,x_l)$ длины $l$, которая записана в виде матрицы размера $l\times n$ ($l$ строк, $n$ столбцов), а целевые значения $y$ множества $Y$ представлены конечным множеством {0, 1, 2}, где 0 обозначает низкий уровень агрессии или ее отсутствие, 1 — средний уровень агрессии, а 2 — высокий уровень агрессии.

Схема предложенного метода для определения депрессии по речи представлены на рис. 2.

Предложенный метод на основе нескольких наборов акустических признаков и ансамбля из методов случайного леса представляет собой решение задачи мультиклассовой классификации. Данные размечены согласно трем уровням уровня агрессии: низкий, средний и высокий. Из аудиоданных были вычислены наборы акустических признаков ComParE 2013, DenseNet и auDeep, после чего была проведена нормализация данных, а затем был применен ансамбль из методов случайного леса различными значениями весов для классов для классификации с голосованием по большинству.

## 2. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ С ИСПОЛЬЗОВАНИЕМ ПРЕДЛОЖЕННОГО МЕТОДА ОПРЕДЕЛЕНИЯ АГРЕССИИ

Корпус SD содержит видеозаписи взаимодействий человек-человек в информационно-справочном центре. Для записи использо-

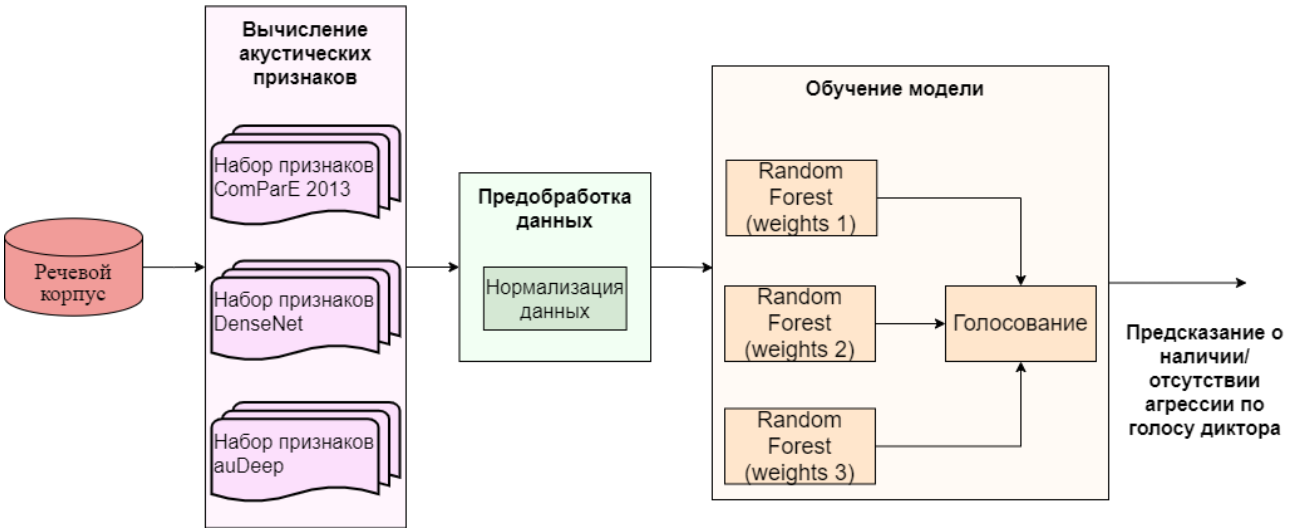


*Рис. 2. Схема предложенного метода определения агрессии в разговорной речи*
*[Fig. 2. Scheme of the proposed method for aggression detection in colloquial speech]*

валось 4 кратких описания ситуаций, которые должны были вызвать стресс у испытуемых. В качестве испытуемых была выбрана группа из 9 актеров разных культурных групп (5 женщин и 4 мужчины), которые были разделены на две группы, а каждый сценарий воспроизводился дважды. Актеры были носителями нидерландского и английского языков. Корпус TR состоит из 21 краткого описания ситуаций нежелательного поведения в поездах и на станциях (например, харассмент, воровство, проезд без билета), которые были сыграны 13 актерами. Актеры были носителями английского языка. Для обучения и тестирования моделей использовались корпуса SD и TR (их основные параметры представлены в табл. 1).

В результате вычисления набора акустических признаков для 293 объекта для обучения и 117 объектов для тестирования было получено 6373, 4097 и 1024 признака для наборов openSMILE, auDeep и DenseNet соответственно. Для классификации при разработке предложенного метода был выбран метод случайного леса с параметрами, подобранными при помощи поиска по сетке. Экспериментальные исследования проводились с использованием техники 5-кратной перекрестной валидации. В качестве показателя качества распознавания был выбран показатель UAR.

Наиболее распространенным показателем качества распознавания для задачи определения агрессии в речи является невзвешенная средняя полнота (Unweighted Average Recall, UAR) — это показатель на основе средней чувствительности и специфичности (mean of sensitivity and specificity), где $N_c^{(i)}$ описывает количество верно распознанных элементов $i$-го класса, $N_0^{(i)}$ описывает общее количество объектов в $i$-м классе, $N$ описывает общее количество объектов, а $k$ — количество классов:

$$UAR = \frac{1}{k}\sum_{i=1}^{k}\frac{N_c^{(i)}}{N_0^{(i)}}.$$

Результаты сравнения с лучшими известными аналогами представлено в табл. 2. Результаты сравнительных экспериментальных исследований представлены в табл. 3. В таблице градиентом от оттенков красного до оттенков зеленого отмечены худшие и лучшие результаты соответственно.

На основе сравнительных экспериментальных исследований можно сделать вывод о том, что лучшие и наиболее стабильные (без резких перепадов в результатах классификации по различным наборам признаков) результаты среди нескольких наборов акустических признаков были получены с использованием метода случайного леса. Лучший результат распознавания агрессии с использованием предложенного метода достиг показателя UAR = 76,5 %. Можно заметить, что предложенный метод является конкурентноспособным в мультиклассовой задаче автоматического определения агрессии по речевым высказываниям.

*Таблица 1. Параметры корпусов SD и TR*
*[Table 1. SD and TR corpora parameters]*

| | Обучающая часть | Настроечная часть | Тестовая часть | Всего |
|---|---|---|---|---|
| Количество записей (низкий уровень агрессии, средний уровень агрессии, высокий уровень агрессии) | 293 (156, 74, 63) | 117 (69, 33, 15) | 501 | 893 |
| Средняя продолжительность записи (сек) | 5 | | | |
| Общая длительность записей (мин) | 75:23 | | | |

*Таблица 2. Результаты сравнения предложенного метода с альтернативными методами [Table 2. Results of comparison of the proposed method with alternative methods]*

| Работа | Данные, модальность | Результат, UAR |
|---|---|---|
| Базовый метод на соревнованиях ComParE-2021 на основе мешка аудио слов и метода опорных векторов [12] | SD, TR, аудио | 72,2 % |
| Метод на основе X-векторов и векторов Фишера и метода опорных векторов [6] | | 77,8 % |
| Предложенный метод для определения агрессии в речевых высказываниях | | 76,5 % |

*Таблица 3. Сравнительные экспериментальные исследования методов в задаче автоматического определения агрессии [Table 3. Comparative experimental studies of methods in the task of automatic aggression detection]*

| | openSMILE, UAR, % | auDeep, UAR, % | DeepSpectrum, UAR, % |
|---|---|---|---|
| Градиентный бустинг (Catboost) | 68,3 | 45,6 | 61,0 |
| Случайный лес (Random Forest) | 70,0 | 50,1 | 65,3 |
| Метод k-ближайших соседей (k-Nearest Neighbours, k-NN) | 37,1 | 33,3 | 52,8 |
| Бэггинг (Bagging) | 63,6 | 48,2 | 53,2 |
| Бэггинг + Метод k-ближайших соседей | 34,0 | 44,0 | 49,2 |
| Бэггинг + Случайный лес | 75,0 | 34,9 | 59,3 |
| Дерево решений (DecisionTreeClassifier) | 60,0 | 47,0 | 44,9 |
| Метод опорных векторов с параметром контроля количества опорных векторов (NuSVC) | 32,7 | 43,0 | 65,3 |
| Метод опорных векторов с линейной функцией ядра (LinearSVC) | 33,3 | 38,0 | 45,6 |
| Мультиклассовая стратегия на основе выходного кода с исправлением ошибок + Метод опорных векторов с линейной функцией ядра (OutputCodeClassifier + LinearSVC) | 33,3 | 38,0 | 55,4 |
| Линейный дискриминантный анализ (Linear Discriminant Analysis, LDA) | 63,6 | 38,0 | 62,6 |
| Мультиклассовая стратегия «один против одного» (OneVsOneClassifier) | 31,3 | 35,0 | 53,9 |
| Мультиклассовая стратегия «один против остальных» (OneVsRestClassifier) | 33,6 | 39,0 | 46,1 |

## ЗАКЛЮЧЕНИЕ

В статье рассматривается задача автоматического определения агрессии в разговорной речи. В работе в кратком виде проанализированы существующие подходы. На основе аналитического обзора и сравнительных экспериментальных исследований был разработан предложенный метод автоматического определения агрессии в разговорной речи на основе ансамбля случайных лесов, обученного на различных наборах акустических при-

знаков с различными весами. С использованием предложенного метода удалось достичь результатов распознавания агрессии 76,5 % по показателю UAR.

Дальнейшие работы связаны с улучшением результатов распознавания путем применения более сложных методов машинного обучения, в том числе, нейронных архитектур. Кроме того, планируются комплексные работы по обработке акустических признаков, включающие как методы аугментации, так и методы уменьшения размерности признакового пространства.

## БЛАГОДАРНОСТИ

Исследование выполнено при финансовой поддержке РНФ в рамках научного проекта № 22-11-00321 (разработка предложенного в статье метода автоматического определения агрессии в разговорной речи) и РФФИ в рамках научного проекта № 20-37-90144.

## КОНФЛИКТ ИНТЕРЕСОВ

Автор декларирует отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

## СПИСОК ЛИТЕРАТУРЫ

1. *Карпов, А. А.* Актуальные задачи и достижения систем паралингвистического анализа речи / А. А. Карпов, Х. Кайа, А. А. Салах // Научно-технический вестник информационных технологий, механики и оптики. – 2016. – Т. 16, № 4. – С.581–592. doi:10.17586/2226-1494-2016-16-4-581-592.

2. *Buss, A.* An inventory for assessing different kinds of hostility / A. Buss, A. Durkee // Journal of Consulting Psychology. – 1957. – Vol. 21(4). – P. 343–349. https://doi.org/10.1037/h0046900.

3. *Busso, C.* IEMOCAP: interactive emotional dyadic motion capture database / C. Busso [et al.] // Language Resour Evaluat. – 2008. – Vol. 42(4). – P. 335-359.

4. *Egas-López, J. V.* Identifying Conflict Escalation and Primates by Using Ensemble X-Vectors and Fisher Vector Features / J.V. Egas-López [et al.] // In Proc. of INTERSPEECH-2021. – 2021. – P. 476–480. doi: 10.21437/Interspeech.2021-1173.

5. *Lefter, I.* An audio-visual dataset of human–human interactions in stressful situations / I. Lefter, G. J. Burghouts, L. J. Rothkrantz // Journal on Multimodal User Interfaces. – 2014. – Vol. 8(1). – P. 29–41.

6. *Lefter, I.* Aggression recognition using overlapping speech / I. Lefter, C. M. Jonker // Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). – 2017. – 299-304. doi: 10.1109/ACII.2017.8273616.

7. *Lefter, I.* A comparative study on automatic audio–visual fusion for aggression detection using meta-information / I. Lefter, L. Rothkrantz, G. Burghouts // Pattern Recognition Letters. – 2013. – Vol. 34(15). – P. 1953–1963.

8. *Livingstone, S. R.* The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English / S. R. Livingstone, F. A. Russo // PLoS ONE. – 2018. – Vol. 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391. – (дата обращения: 01.07.2022).

9. *McWilliams, N.* Psychoanalytic diagnosis: Understanding personality structure in the clinical process / N. McWilliams 2nd ed. Guilford Press. – 2011. – 426 с.

10. *Perepelkina, O.* RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition / O. Perepelkina, E. Kazimirova, M. Konstantinova // PeerJ Preprints 6:e26688v1. – 2018. https://doi.org/10.7287/peerj.preprints.26688v1. – (дата обращения: 01.07.2022).

11. *Sahoo, S.* Detecting Aggression in Voice Using Inverse Filtered Speech Features / S. Sahoo, A. Routray // IEEE Transactions on Affective Computing. – 2018. – Vol. 9(2). DOI: 10.1109/TAFFC.2016.2615607.

12. *Schuller, B.* The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates / B. Schuller [и др.] // In Proc. of INTERSPEECH-2021. – 2021. – P. 431-435. doi: 10.21437/Interspeech.2021-19.

13. *Sobin, C.* Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy / C. Sobin, M. Alpert // J Psycholinguist. – 1999. – Vol. 28. – P. 347–365. https://doi.org/10.1023/A:1023237014909.

14. *Zadeh, A.* Multi-attention recurrent network for human communication comprehension / A. Zadeh [et al.] // Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. – 2018. – С. 5642-5649.

15. *Zhou, Z.* Detecting Escalation Level from Speech with Transfer Learning and Acoustic-Lexical Information Fusion / Z. Zhou, Y. Xu, M. Li // arXiv preprint. arXiv:2104.06004v2.

**Величко Алёна Николаевна** — н.с. лаборатории Речевых и многомодальных ин¬терфейсов Санкт-Петербургского Федерального исследовательского центра РАН.
E-mail: alena.n.velichko@gmail.com
ORCID iD: https://orcid.org/0000-0002-8503-8512

DOI: https://doi.org/10.17308/sait/1995-5499/2022/4/180-188 ISSN 1995-5499
Received 14.07.2022
Accepted 05.12.2022

# A SPEECH SINGNAL ANALYSIS METHOD FOR AUTOMATIC AGGRESSION DETECTION IN COLLOQUIAL SPEECH

**© 2022 A. N. Velichko**✉

*St. Petersburg Federal Research Center of the Russian Academy of Sciences*
*39, 14 Line V.O., 199178 Saint-Petersburg, Russian Federation*

**Annotation.** In recent years the destructive behaviour detection in the Internet task becomes more popular aiming to provide users' psychological comfort. Destructive behaviour includes aggression which in European culture is presented as motivated destructive behaviour that can be oriented to the outside and the inside. Also, such behaviour contradicts currently accepted social norms. This paper implicates aggression as a paralinguistic phenomenon in the way that it reveals in speech and not what exactly was pronounced. This paper presents the definition and types of aggression as well as a short analysis of existing approaches for aggression detection in colloquial speech. The formalization of the multiclass classification task and description of the proposed approach also were presented in the paper. The experiments were made on the classification methods for automatic aggression detection, where the best result was achieved by the random forest. With the use of the random forest, we have got the best and most stable results. Based on the experiments the proposed approach for aggression detection was developed. Audio files from the mutlimodal corpora Stress at Service Desk Dataset and Aggression in Trains were used to train and test the models with the use of 5-fold cross-validation. The proposed approach includes an ensemble of random forests that were trained on different acoustic feature sets with different weights. The best result achieved using the proposed approach is 76.5 % in terms of unweighted average recall and is one of the best results achieved by other scientific groups.
**Keywords:** computational paralinguistic, destructive behaviour, aggression, automatic aggression detection in colloquial speech, speech technologies, multiclass classification, machine learning.

## CONFLICT OF INTEREST

The author declare the absence of obvious and potential conflicts of interest related to the publication of this article.

✉ Velichko Alena N.
e-mail: alena.n.velichko@gmail.com

## REFERENCES

1. *Karpov A., Kaya H. and Salah A.* (2016) State-of-the-art tasks and achievements of paralinguistic speech analysis systems. *Scientific and Technical Journal of Information Technologies Mechanics and Optics*. 16(4). P. 581–592. doi:10.17586/2226-1494-2016-16-4-581-592. (in Russian)

2. *Buss A. and Durkee A.* (1957) An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*. 21(4). P. 343–349. https://doi.org/10.1037/h0046900.

3. *Busso C., Bulut M., Lee C. [et al.]* IEMOCAP: interactive emotional dyadic motion capture database. *Language Resour Evaluat*. 2008. 42(4). P. 335–359.

4. *Egas-López J. V., Vetráb M., Tóth L. and Gosztolya G.* (2021) Identifying Conflict Escalation and Primates by Using Ensemble X-Vectors and Fisher Vector Features. I*n Proc. of INTERSPEECH-2021*. P. 476–480. doi: 10.21437/Interspeech.2021-1173.

5. *Lefter I., Burghouts G. J. and Rothkrantz L. J.* (2014) An audio-visual dataset of human–human interactions in stressful situations. *Journal on Multimodal User Interfaces*. 8(1). P. 29–41.

6. *Lefter I. and Jonker C.* M. (2017) Aggression recognition using overlapping speech. *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. P. 299–304. doi: 10.1109/ACII.2017.8273616.

7. *Lefter I., Rothkrantz L. and Burghouts G.* (2013) A comparative study on automatic audio–visual fusion for aggression detection using meta-information. *Pattern Recognition Letters*. 34(15). P. 1953–1963.

8. *Livingstone S. R. and Russo F. A.* (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE*. 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

9. *McWilliams N.* (2011) Psychoanalytic diagnosis: Understanding personality structure in the clinical process. 2nd ed. *Guilford Press.* 426.

10. *Perepelkina O., Kazimirova E. and Konstantinova M.* (2018) RAMAS: Russian Multimodal Corpus of Dyadic Interaction for studying emotion recognition. *PeerJ Preprints* 6:e26688v1. https://doi.org/10.7287/peerj.preprints.26688v1.

11. *Sahoo S. and Routray A.* (2018) Detecting Aggression in Voice Using Inverse Filtered Speech Features. I*EEE Transactions on Affective Computing*. 9(2). DOI: 10.1109/TAFFC.2016.2615607.

12. *Schuller B., Batliner A., Bergler C. [et al.]* (2021) The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. *In Proc. of INTERSPEECH-2021*. P. 431–435. doi: 10.21437/Interspeech.2021-19.

13. *Sobin C. and Alpert M.* (1999) Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy. *J Psycholinguist.* 28. 347–365. https://doi.org/10.1023/A:1023237014909.

14. *Zadeh A., Liang P., Poria S. [et al.]* (2018) Multi-attention recurrent network for human communication comprehension. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. P. 5642–5649.

15. *Zhou Z., Xu Y. and Li M.* (2021) Detecting Escalation Level from Speech with Transfer Learning and Acoustic-Lexical Information Fusion. *arXiv preprint*. arXiv:2104.06004v2.

**Velichko Alena N.** — researcher of the Speech and Multimodal Interfaces Labo¬ratory of St. Petersburg Federal Research Center of the Russian Academy of Sciences.
E-mail: alena.n.velichko@gmail.com
ORCID iD: https://orcid.org/0000-0002-8503-8512