УДК 004.934.2 **DOI** 10.15622/ia.23.1.1

А.А. Поволоцкая, А.А. Карпов

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ АВТОМАТИЧЕСКОГО АНАЛИЗА ЭКСТРАЛИНГВИСТИЧЕСКИХ КОМПОНЕНТОВ СПОНТАННОЙ РЕЧИ

Поволоцкая А.А., Карпов А.А. **Аналитический обзор методов автоматического анализа экстралингвистических компонентов спонтанной речи.**

Аннотация. Точность систем автоматического распознавания спонтанной речи далека от тех, которые демонстрируют системы распознавания подготовленной речи. Обусловлено это тем, что спонтанная речь не характеризуется той плавностью и отсутствием сбоев, что подготовленная. Спонтанная речь варьируется от диктора к диктору: отличное произношение фонем, наличие пауз, речевых сбоев и экстралингвистических компонентов (смех, кашель, чихание, и цыканье при выражении эмоции раздражения и др.) прерывают плавность вербальной речи. Экстралингвистические компоненты очень часто несут важную паралингвистическую информацию, поэтому для систем автоматического распознавания спонтанной речи важно распознавать подобные явления в потоке речи. В данном обзоре проанализированы научные работы, посвященные проблеме автоматического анализа экстралингвистических компонентов спонтанной речи. Рассмотрены и описаны как отдельные методы и подходы по распознаванию экстралингвистических компонентов в потоке речи, так и работы, связанные с многоклассовой классификацией изолированно записанных экстралингвистических компонентов. Наиболее распространенными методами анализа экстралингвистических компонентов являются нейронные сети, такие как глубокие нейронные сети и сети на основе моделей-трансформеров. Приведены основные понятия, относящиеся к термину экстралингвистические компоненты, предложена оригинальная систематизация экстралингвистических компонентов в русском языке, описаны корпуса и базы данных звучащей разговорной речи как на русском, так и на других языках, также приведены наборы данных экстралингвистических компонентов, записанных изолированно. Точность распознавания экстралингвистических компонентов повышается при соблюдении следующих условия работы с речевым сигналом: предобработка аудиосигналов вокализаций показала повышение точности классификации отдельно записанных экстралингвистических компонентов; учет контекста (анализ нескольких фреймов речевого сигнала) и использовании фильтров для сглаживания временных рядов после извлечения векторов признаков показали повышение точности при пофреймовом анализе речевого сигнала со спонтанной речью.

Ключевые слова: автоматическое распознавание речи, речевые технологии, машинное обучение, прикладная лингвистика, экстралингвистические компоненты, спонтанная речь, автоматическое распознавание экстралингвистических компонентов.

1. Введение. Поток вербальной речи характеризуется наличием паралингвистических и экстралингвистических средств, которые, формируя речевой портрет человека, способствуют распознаванию его психоэмоционального и физического состояний. Паралингвистические фонационные явления, такие как интонация, мелодика, длина пауз, темп, тембр, громкость речи, часто используются при решении задач

классификации эмоций. В то время как, экстралингвистические средства включают в себя паузы, и различные психоэмоциональные проявления невербального характера (плач, смех, кашель, вздохи/выдохи и др.). Несмотря на достигнутый существенный прогресс в области автоматического распознавания речи (АРР), до сих пор уделяется недостаточно внимания анализу экстралингвистических средств речи. Анализ экстралингвистических средств позволит решить сложные научно-технические вопросы в области АРР, распознавания психоэмоциональных и физических состояний человека, поскольку, подобные невербальные явления могут содержать больше информации, чем вербальное выражения эмоций.

Экстралингвистические средства часто расцениваются, как «зашумление» исходного речевого сигнала или «артефакты», которые, присутствуя в потоке речи, нарушают работу систем АРР или запросно-ответных диалоговых систем [1]. Также, определенную сложность создают явления, воспроизводимые совместно с речью, например, когда человек смеется и пытается одновременно что-то сказать (речевой смех). С другой стороны, на данный момент нет единого мнения и определенного термина, который бы характеризовал экстралингвистические средства, присутствующие в речи человека.

Совершенствование методов анализа экстралингвистических средств позволит применять их в различных областях, которые перечислены ниже.

1) Биометрические технологии: распознавание и верификация диктора [2, 3].

2) Автоматическая расширенная транскрипция [4, 5].

3) Учет экстралингвистических средств при распознавании эмоций в речи [4, 6].

4) Оценка здоровья говорящего: детектирование кашля, как симптом заболевания дыхательных путей или проявления психоэмоционального состояния [7, 8].

5) Усовершенствование устройств для контроля родителями новорожденных или маленьких детей, оставшихся в комнате одни (радио няня), путем классификации детских вокализаций: плач, смех, маргинальный/канонический лепет [9].

6) Совершенствование диалоговых систем и голосовых ассистентов, как при распознавании речи пользователя, так и при составлении ответа системы [10, 11]. Возможность «очеловечить» голосового ассистента и диалоговые системы путем включения экстралингвистических средств в поток речи, синтезируемой системой.

7) Совершенствование человеко-машинного взаимодействия для людей с ограниченными возможностями, т.е. управление мобильными устройствами и компьютерами с помощью определенных команд-звуков, например, щелканье языком – «выбрать элемент», более интенсивный выдох – «отправить сообщение» и т.д. [12].

Данная статья посвящена обзору современных технологий автоматического анализа экстралингвистических средств. В разделе 2 содержатся основные понятия и оригинальная систематизация. В разделе 3 приведен сравнительно-сопоставительный обзор существующих корпусов и наборов данных, на основе которых возможно обучение моделей систем для анализа экстралингвистических средств в спонтанной речи. На основе проведенного обзора предложена методика записи корпуса со спонтанной речью. В разделах 4 и 5 представлен аналитический обзор подходов и методов автоматического анализа экстралингвистических средств как в спонтанной речи (СР), так и распознавание экстралингвистических средств, записанных изолировано. Выводы по проделанной работе оформлены в разделе 6.

2. Основные понятия и систематизация экстралингвистических компонентов. В [13] явления, подобные смеху, кашлю, икоте и др. называются невербальные вокализации (НВ, *nonverbal vocalization*). В [14] подобные краткие, невербальные произнесения, возникающие между эпизодами речи называются вокальные всплески (*vocal bursts*). Одна и та же НВ может представлять реализацию различных явлений. Например, вздох может являться как вегетативным звуком (процесс дыхания), так и быть частью смеха, или быть прагматическим сигналом со значением раздражения, усталости или просьбы обратить внимание.

В [15] предложено разделение НВ на четыре следующих уровня:

1) вегетативный уровень – явления, подобные дыханию, чиханию, и другим рефлекторным реакциям, которые могут сигнализировать о состоянии здоровья человека, но и использоваться для распознавания личности говорящего.

2) орфографический уровень – представляет НВ, которые могут быть записана в виде последовательности звуков (фонем) одного языка и быть понятной для носителей этого языка (орфографическая невербальная вокализация).

3) аффективный уровень – аффективные звуки и большинство междометий определяют аффективное и психоэмоциональное состояние человека. Аффективная информация

обычно не присутствует в вегетативных звуках и звуках-заполнителях. Звуки обратной связи (ответные лексемы), однако, иногда могут передавать аффективную информацию.

4) прагматический уровень – представляет выражение некоторых НВ в роли прагматических частиц, которые выполняют функцию управления разговором. Например, звуки обратной связи (ответные лексемы), смех и аффективные звуки могут выполнять функцию поддержания разговора.

В результате проведенного обзора литературы было решено: систематизация экстралингвистических средств и их объединение под общим термином вокализации не представляется возможным, поскольку, в процессе вокализации участвуют голосовые связки. В связи с чем решено разделить экстралингвистические средства на экстралингвистические явления (ЭЯ), т.е. кашель, чихание, цыканье и другие звуки, при воспроизведении которых голосовые связки не принимают участие, и экстралингвистические вокализации (ЭВ), когда голосовые связки вибрируют, и объединить экстралингвистические явления и вокализации под общим термином экстралингвистические компоненты (ЭК) вербальной речи.

На рисунке 1 представлена оригинальная систематизация ЭК для русского языка. Все ЭК можно разделить на: 1) врожденные, 2) приобретенные.

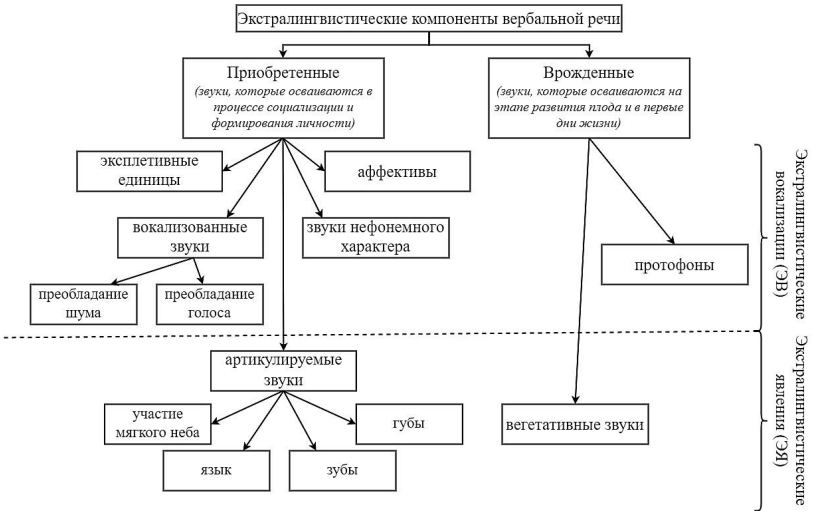

Рис. 1. Систематизация экстралингвистических компонентов в русском языке

1) Врожденные ЭК. Подразделяются на вегетативные (рефлекторные) и протофоны, т.е. звуки, которые закладываются у человека во время внутриутробного развития и формируются во время первых дней жизни:

а. вегетативные (рефлекторные) звуки такие как: вдох-выдох, глотательные звуки, жевательные звуки, одышка (после физической нагрузки), в первую очередь, не являются коммуникативными сигналами и не все находятся под осознанным контролем. Как правило, вегетативные звуки не «изучаются» и не «заучиваются» человеком, а являются естественными физиологическими процессами, контролируемыми подсознательно. Также в данную группу входят естественные реакции организма человека на определенные возбудители/раздражители: кашель, икота, зевота, чихание, отрыжка и др. В данную группу также возможно отнести плач и крик (у детей это сигналы дистресса), смех, и другие эмоциональные реакции, провоцируемые внешними и внутренними (психоэмоциональные процессы) стимулами.

б. к вокализациям у младенцев также относятся протофоны (речеподобные звуки), которые могут быть свободными вокализациями – отдельные звуки, квазисогласные, маргинальный и канонический лепет. Воспроизводство протофонов является эндогенным и сигнализирует о психофизиологическом состоянии ребенка на довербальной стадии развития, в то время как крик, смех, плач несут социальную функцию и выражают определенные эмоциональные состояния [16].

2) Приобретенные ЭК. В раннем возрасте звуки дистресса (смех, плач, крик, и др.) являются непроизвольными и не контролируются ребенком осознанно. По мере взросления, человек овладевает навыками коммуникации, взаимодействуя с другими членами общества, осваивает правила и нормы коммуникации, и «научается» контролировать различные проявления своих эмоций в соответствии с культурными особенностями общества.

а. звуки, артикулируемые с помощью речевого аппарата, можно разделить по активному действующему органу: язык (щелканье, цыканье, др.); губы (причмокивание, свист, др.); зубы (стук, скрипение, др.); звуки, реализуемые посредством мягкого неба, а именно, при его опускании (звуки храпа, хрюканья и др.).

б. вокализованные звуки (или аффективные звуки) могут реализовываться с преобладанием голоса (во время фонации голосовые связки будут сведены, напряжены и находиться в состоянии колебания) в таком случае воспроизводится крик, возгласы,

плач, смех, и др.; так и шума (нет смычки между голосовыми связками и отсутствует колебание) может быть реализовано что-то похожее на кряхтение, шипение, кашель и др.

в. звуки нефонемного характера реализуются во время заполненных пауз хезитации: гласные без цели, плюс потенциальный нейтральный носовой согласный (*«эм»* или *«ммм»*); гудящие сигналы, (*«хм»*, *«ага»*, *«угу»* и др. – обычно они используются как сигналы обратного ответа, но потенциально используются также для выражения отношения в разговоре); воспроизведение мелодий без текста с использованием одного звука («промычать» или «промурлыкать»).

г. эксплетивные единицы – десемантизированные элементы разговорной речи, которые являются, скорее, формой удержания канала связи, к ним также относятся дискурсивные слова (ДС): *«ну» «это»*, *«которое», «типа»*, *«во-о-от»*, и др.

д. аффективы или междометия не имеют адресата сообщения и являются формой выражения эмоции различной степени лексикализации: «*ба!*», «*ого!*», «*ой*», «*тс-с-с-с*» и др.

Речь каждого индивида уникальна и может характеризоваться индивидуальными особенностями, как врожденными, так и приобретенными:

а. возможно частое употребление ДС. Что может быть проявлением речевой привычки, заполнителем пауз хезитации, выражением эмоционального состояния или мнения, маркерами членения текста, способом управления вниманием слушающего.

б. наличие синдромов речевых нарушений, которые могут быть как врожденными (наследственные заболевания с умственной отсталостью, характеризующиеся проблемами развития речи: затяжки звуков и/или их частое повторение), и приобретенными, возникающие при органических расстройствах речевого отдела коры головного мозга в результате перенесенных травм, опухолей, инсультов и др.

в. наличие нейродегенеративных болезней: боковой амиотрофический склероз, болезни Альцгеймера и Паркинсона и др., когда наблюдается заметное изменение в порождении речи.

Такую систематизацию можно признать универсальной, поскольку она описывает и объединяет все ЭК по характеристики врожденности-приобретенности и не рассматривает ЭК с точки зрения прагматической составляющей, т.е. мотивированности использования в потоке речи.

3. Базы данных звучащей разговорной речи и отдельно записанных экстралингвистических компонентов. Наиболее часто ЭК встречаются в СР. СР – неподготовленная и экспромтная,

ситуативно обусловленная речевая деятельность, используемая в контексте повседневного неформального общения между собеседниками, которые состоят в неофициальных отношениях, включающая элементы нормированного и ненормированного произношения.

Систематизированные данные обзора корпусов СР представлены в таблицах 1 (русская речь) и 2 (речь на иностранных языках), в которых содержится следующая информация по корпусам: общая длительность звучания в часах (Д); количество дикторов их пол (М – мужчины, Ж – женщины) и возраст (В); методика записи корпуса; уровни разметки: орфографическая расшифровка (О), хезитации и речевые сбои (Х), экстралингвистические компоненты (ЭК), («+» – выполнена полная разметка, «±» – выполнена частичная разметка, «–» – разметка отсутствует); режим доступа.

Таблица 1. Сравнение корпусов спонтанной русской речи

Корпус	Дикторы			Методика записи	Д	Уровни разметки			Доступ
	В	М	Ж			О	Х	ЭК	
CAT [17]	18–52	89	123	чтение и пересказ текста, описание изобр., рассказ на тему	~50	–	–	–	По запросу
ОРД [18]	18–83	69	61	повседневное общение	~5	+	+	+	Закрытый
Корпуса звучащей речи [19]	7–17	59	67	рассказ о сновидениях	~5	+	+	±	По запросу[1]
	19–70	8	9	рассказы сибиряков о жизни					
	20–30	12	28	рассказы из жизни					
CORUSS [20]	16–30	10	10	чтение текста, монолог, диалог	>10	–	+	+	По запросу
	31–45	10	10						
	46–77	10	10						
РМД [21]	18–36	34	62	пересказ фильма – полилог 3 участников	~17	+	+	+	По запросу[2]

Проанализированы также корпуса, содержащие речь на других языках. В таблице 2 представлены не только американский (The Buckeye speech corpus [22]) и британский английский (LUCID [23], SSPNet Vocalization Corpus [24]), но и интерферированная речь китайцев: мандаринский, сингапурский и тайванский диалекты китайского (The Wildcat Corpus [25], SpiCE [26]), речь корейцев,

[1] http://spokencorpora.ru/
[2] https://multidiscourse.ru/main/

испанцев, турков и дикторов других национальностей (ALLSSTAR [27], The ICSI Meeting Corpus [28]). Также, в таблице 2 приведена информация по корпусам на других языках: китайском [28], венгерском [30], которые часто встречаются в исследованиях по распознаванию СР и ЭК.

Таблица 2. Сравнение корпусов разговорной речи на иностранных языках

Корпус	Дикторы			Методика записи	Д	Уровни разметки			Доступ
	В	М	Ж			О	Х	ЭК	
Английский язык									
Buckeye [22]	18–30	10	10	интервью	~19	+	+	+	Открытый[3]
	40–60	10	10						
LUCID [23]	19–29	20	20	поиск отличий на изобр. в парах	~73	+	±	±	Открытый[4]
SSPNet [24]	18–64	57	63	сравнение ответов на вопросы	~9	+	+	+	По запросу
Wildcat [25]	18–33	20	20	поиск отличий на изобр. в парах	~40	+	+	+	Открытый[5]
	22–34	36	24						
SpiCE [26]	19–34	10	17	рассказ по ряду изобр., интервью	~10	+	±	±	Открытый[6]
ALLS-STAR [27]	18–41	86	54	параллельный перевод; история и поиск отличий по изобр.; интервью	~48	+	±	±	Открытый[7]
The ICSI Meeting Corpus [28]	20–62	40	13	запись полилогов в конференц-зале Международного института компьютерных наук	~72	+	+	+	Открытый[8]
Китайский язык									
NNIME [29]	19–30	20	22	спонтанные пьесы между двумя актерами	~11	+	+	+	Открытый[9]
Венгерский язык									
BEA HSLD [30]	20–90	112	168	рассказ и выражение мнения; полилог по теме; пересказ текста	~260	+	+	±	Закрытый

[3] https://buckeyecorpus.osu.edu/
[4] https://valeriehazan.com/wp/index.php/lucid-corpus-london-ucl-clear-speech-in-interaction/
[5] https://groups.linguistics.northwestern.edu/speech_comm_group/wildcat/
[6] https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP2/MJOXP3
[7] https://speechbox.linguistics.northwestern.edu/ALLSSTARcentral/#!/recordings
[8] https://groups.inf.ed.ac.uk/ami/icsi/
[9] https://nnime.ee.nthu.edu.tw/

Можно выделить несколько основных методик сбора данных со СР: интервью или ответы на вопросы, с которыми диктор не был ознакомлен до проведения записи; рассказ или монолог на определенную тему (чаще всего о пережитом опыте диктора); описание картинок или составление рассказа по ряду изображений; карта или поиск маршрута.

Если для участия в записи разговорной спонтанной речи привлекать пары дикторов (друзей, влюбленных, родителей с детьми, и других людей), которые продолжительное время состоят в близком неформальном контакте, то можно создать более неформальную обстановку в процессе общения в рамках студийной записи, чем если будут приглашены два абсолютно незнакомых человека, которым будет выдано задание.

В качестве методики сбора корпуса со СР предлагается использовать дикторов, состоящий в близком контакте продолжительное время и представлять им игровой формат коммуникации, в которой дикторам не указывается конкретная цель (проложить маршрут на карте, найти отличия и др.), но где необходимо использовать ассоциативное мышление, также, стоит предположить, что, если ограничить количество каналов информации (например, оставить только речевой), то возможно получить более вариативную речь, с большим количеством озвученных ЭК.

Для анализа ЭК в рамках русской СР целесообразно использование следующих корпусов: «CAT» [17], «Корпуса звучащей речи» [19], «CORUSS» [20], поскольку в этих корпусах представлены записи звучащей русской речи, наиболее близкой к спонтанной, также, представлена речь носителей русского языка, в дальнейшем возможна апробация системы анализа ЭК на других языках.

Проведен обзор наборов данных с ЭК, записанными отдельно от речи. Сравнение наборов данных представлено в таблице 3, где содержится информация о продолжительности звучания (количество данных: Ч – часы, Ф – количество фрагментов); список реализованных ЭК и их количество – ЭК (количество); режиме доступа.

Для изучения ЭК и последующего обучения моделей, или отдельных блоков ИНН необходимо использование наборов данных с ЭК, реализованными изолированно: «Multiclass cough sound dataset» [7], «ASVP-ESD» [31], FSD50K (Human Sounds) [34], «VocalSound» [35], поскольку, данные корпуса находятся в открытом доступе и, в общей сложности, имеют широкий спектр реализованных ЭК.

Таблица 3. Наборы данных с экстралингвистическими компонентами

Набор данных	Объем данных		ЭК (количество)	Доступ
	Ч	Ф		
Multiclass cough sound [7]	~1	2944	кашель: бронхиальная астма (787); covid-19 (907); сердечная недостаточность (554); здоровое дыхание (696)	Открытый[10]
ASVP-ESD [31]	~11	3 104	плач (892); смех (772); крик (679)	Открытый[11]
H–VB [32]	~37	59 201	вздох; смех; плач; крик; ворчание/хрюканье; стон; тяжелое дыхание; др.	Закрытый[12]
MAHNOB [33]	~4	1 697	смех (563); речевые высказывания (849); удивление (494); постановочный смех (51); речевой смех (67); другие вокализации (167)	По запросу[13]
FSD50K (Human Sounds) [34]	~23	20 004	возглас (1368); крик (1246); смех (5696); плач/всхлипывание (1462); вопль/стон (215); вздох (321); завывания (702); ворчание (334); свист (1985); дыхание (834); кашель (871); чихание (1200); внюхивание (205); жевание (829); кусание (366); полоскание (137); отрыжка (1302); икота (931); др.	Открытый[14]
VocalSound [35]	~24	21 024	смех; вздох; кашель; прочищение горла; чихание; вдыхание/сопение (все звуки по 3504 фрагментов)	Открытый[15]

Не стоит исключать возможного увеличения наборов данных путем либо накладывания звуковых сигналов с ЭК на вербальную речь, либо осуществлять поиск подобных пересечений в открытых источниках, связано это с тем, что во время речепорождения

[10] http://web.firat.edu.tr/turkertuncer/acute_asthma_cough.rar.
[11] https://www.kaggle.com/datasets/dejolilandry/asvpesdspeech-nonspeech-emotional-utterances
[12] https://doi.org/10.5281/zenodo.6320973
[13] https://mahnob-db.eu/laughter/
[14] https://zenodo.org/record/4060432
[15] https://github.com/YuanGongND/vocalsound

возможны такие явления как речевой смех, или, когда во время плача/приступа икоты человек пытается что-то произнести.

4. Методы автоматического анализа ЭК, записанных отдельно от речи, в рамках соревнований Affective Vocal Bursts (A-VB'22). Известно не так много исследований, посвященных анализу ЭК в СР. Исследования ведутся либо по анализу одной-двух определенных реализаций ЭК во всем речевом сигнале, либо осуществляется классификация фрагментов ЭК, записанных отдельно от речи.

В 2022 году на международной конференции ACII (Affective Computing + Intelligent Interaction) состоялся семинар в рамках соревнований, посвященных распознаванию аффективных вокальных всплесков (Affective Vocal Bursts, A-VB[16]), записанных отдельно от речи [32]: смех, вздох, ворчание и др. В рамках соревнования был представлен большой набор данных Hume-VB[17], описание которого приведено в таблице 3. Весь набор данных (59 201 аудиофрагмент) был поделен на 3 выборки: обучающая, валидационная и тестовая, в каждой по 19 900 аудиофрагментов. Участники представляли работу с описанием метода, где указывали результаты классификации, полученные на тестовом выборке.

Соревнование включало 4 различных задания:

1) *A-VB-High – Классификация эмоций*. Участникам предстояло построить модель многоцелевой регрессии для распознавания 10 эмоций, связанных с вокальными всплесками. По результатам экспериментов, участники сообщали среднее значение коэффициента корреляции конкордации (Concordance Correlation Coefficient, CCC) по всем десяти эмоциям.

2) *A-VB-Two – Классификация валентности и степени возбуждения эмоции* – регрессионная задача, в которой модель должна определить значения возбуждения и валентности (в соответствии со следующими значениями: 1 – неприятное/подавленное, 5 – нейтральное, 9 – приятное/стимулированное), представленные в циклической модели аффектов. Участники сообщали организаторам среднее значение CCC по двум измерениям.

3) *A-VB-Culture – Классификация эмоций с учетом культурной принадлежности диктора* – многоцелевая регрессионная задача. Модель должна распознать интенсивность 40 эмоций (по 10 для каждого языка/культуры), представленного в наборе данных по

[16] https://www.competitions.hume.ai/avb2022
[17] https://zenodo.org/record/6308780

дикторам из следующих стран: США (английский язык), Китай (мандаринский китайского язык), Южная Африка (английский), Венесуэла (испанский язык). Все участники сообщали среднее значение CCC по всем 40 эмоциям.

4) *A-VB Type – Классификация типа вокализаций*. Модель классифицирует типы экспрессивного вокального всплеска по 8 классам (вздох, смех, плач, крик, ворчание/хрюканье, стон, тяжелое дыхание, другое). Участники сообщали значение показателя невзвешенного среднего (Unweighted Average Recall, UAR) в качестве показателя точности.

В рамках обзора основной фокус смещен в сторону четвертого задания – *A-VB Type – Классификация типа вокализаций*, где основной задачей является классификация ЭК (результаты предложенных моделей представлены в таблице 4).

Таблица 4. Сравнение результатов работы методов на соревновании A-VB'22 в задании *Классификация типа вокализаций* (UAR, %)

Работа	Метод классификации	Результат	
		Валидационная выборка	Тестовая выборка
[37]	data2vec-SM	**58,40**	**58,60**
[38]	wav2vec 2.0	56,86	56,18
[39]	wav2vec2 – XLSR53	49,89	49,70
[32]	End2You (базовый метод)	41,66	41,72

Ниже приведен обзор методов, представленных в соревновании.

4.1. Базовый метод [32]. В качестве базового метода авторами был выбран инструментарий для мультимодального профилирования с помощью сквозного глубокого обучения End2You. Метод приведен на рисунке 2.

Признаки извлекаются покадрово из необработанного сигнала с помощью нейронной сети Emo-18 CNN, предобученную на наборе данных RECOLA [36]. За ней следуют двуслойная сеть долгой краткосрочной памяти (Long Short-Term Memory, LSTM), которая для последующей классификации использует временные паттерны сигнала. Результаты классификации реализованной базовой модели представлены в таблице 4, на валидационной и тестовой выборках результат составил 41,66% и 41,72% по показателю UAR, соответственно.

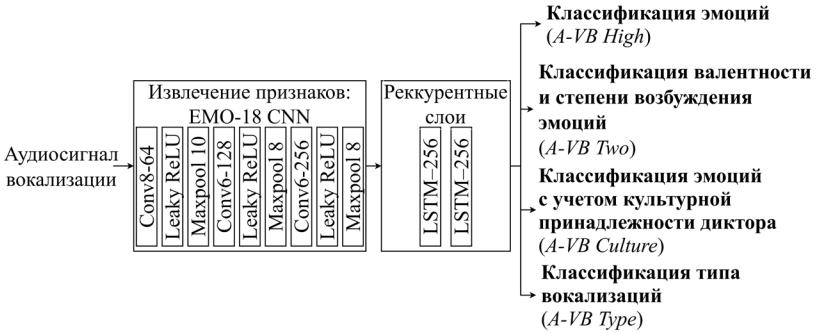

Рис. 2. Схема базового метода в рамках соревнования A-VB'22 [32]

4.2. Метод на основе data2vec-SM [37]. Для классификации ЭК группа авторов использовала предобученную модель data2vec. Схема метода представлена на рисунке 3.

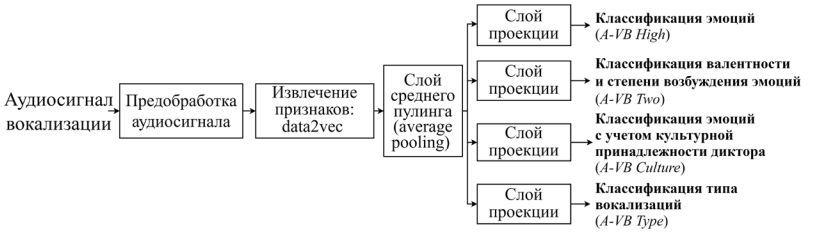

Рис. 3. Схема метода на основе data2vec [37]

Необработанный аудиосигнал подается в предварительно обученную модель data2vec, которая включает блок предобработки. Переменная длина последовательности дополняется нулями до самой длинной последовательности. Из-за маски внимания (attention mask), извлеченные признаки различаются по длине, поэтому они объединяются с помощью операции среднего пулинга (average pooling) и передаются в последующие слои проекции.

Проекционные слои уменьшают размерность выходных данных с помощью функции GELU (Gaussian Error Linear Unit), которая уменьшает размерность до требуемой для конкретной задачи (*Классификация типа вокализации* = 8).

Для классификации используется функция активации Softmax, функция потерь – перекрестная энтропия (cross-entropy, CE) с

использованием обратно пропорциональных весов классов. Затем значения потерь линейно объединяются через найденные оптимальные веса. Результаты представлены в таблице 4, на валидационной и тестовой выборках результат составил 58,40% и 58,60% по показателю UAR, соответственно.

4.3. Метод на основе wav2vec 2.0 [38]. В основу данного метода легло использование большой предобученной модели с режимом самообучения (self-supervised learning) wav2vec2.0. Особенность архитектуры: признаки, полученные с помощью wav2vec2.0, являются отправной точкой для первой задачи (*Классификация типа вокализаций (A-VB Type)*), затем полученные прогнозы для первого задания объединяются с прогнозами каждой последующей задачи, и используются в качестве входных данных для следующей задачи. Метод представлен на рисунке 4.

В качестве адаптивной стратегии инициализации весов использовался метод Dynamic Weight Averaging (DWA), который является частным случаем Softmax распределения, с помощью которого инициализируются веса λ_k для каждого *k*-ого класса [39]:

$$\lambda_k(t) = \frac{K \exp(\frac{L_k(t-1)}{L_k(t-2)} / T)}{\sum_{i=1}^{N} \exp(w_i(t-1)/T)}, \quad (1)$$

где k – класс из множества классов K в рамках многозадачного обучения; t – индекс итерации; T – «температура» Softmax, которая управляет мягкостью распределения вероятностей в задаче; L – значение потери, N – количество состояний для каждого i-го состояния, где каждое i-ое состояние – класс. Большое значение T приводит к более равномерному распределению между различными задачами. Если значение T достаточно велико, тогда $\lambda_i \approx 1$ и задачи взвешиваются одинаково. Оператор Softmax, который умножается на количество классов K гарантирует, что $\sum_{i=1}^{N} \lambda_i(t) = K$. В реализации [38] значение потери $L_k(t)$ рассчитывается как среднее значение потери в каждой эпохе за несколько итераций, что позволяет уменьшить неопределенность, возникающую при стохастическом градиентном спуске и случайном выборе обучающих данных. Две первые итерации t =1 и 2 в [38] были проинициализированы, как $w_k(t) = 1$.

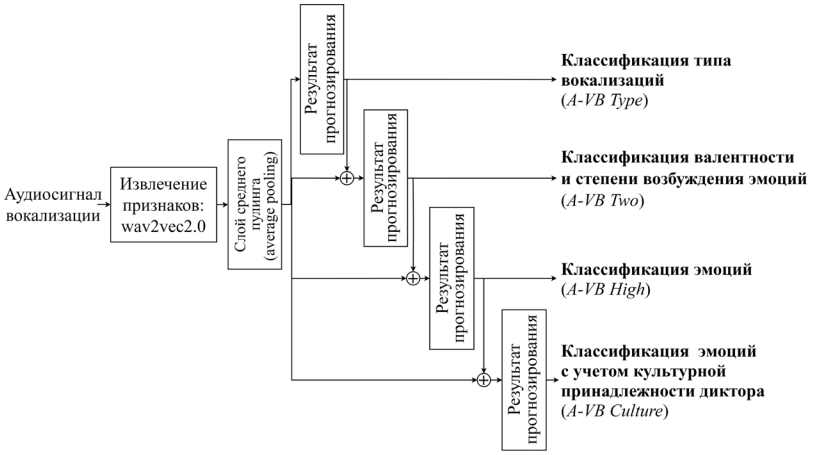

Рис. 4. Схема метода на основе wav2vec 2.0 в рамках соревнования A-VB'22 [38]

Общая сумма потерь становится взвешенной суммой [38]:

$$L_{DWA} = \sum_{k=1}^{K} \lambda_k(t) L_k \,, \tag{2}$$

где λ_k – проинициализированные веса на итерации t, L_k – значение функции потерь.

Результаты представлены в таблице 4, на валидационной и тестовой выборках результат составил 56,86% и 56,18% по показателю UAR, соответственно.

4.4. Метод на основе wav2vec2-XLSR53 [40]. Модель wav2vec-XLSR53 представляет базовую конфигурацию трансформера, обученного на трех наборах неразмеченных данных: Multilingual LibriSpeech [41], Common Voice [42], BABEL [43], которая использовалась для извлечения признаков аудиосигнала. Схема модели представлена на рисунке 5.

Последовательность признаков, полученных из аудиосигнала моделью wav2vec-XLSR53, подается на вход сети LSTM. За слоем LSTM следует полносвязный слой (Fully Connected, FC). Использование FC обосновано необходимостью сокращения размерности вложения до значения, соответствующего значению выходного слоя в зависимости от задания, затем слой последнего пулинга (last pooling) [44], т.е. используется последний вектор

признаков из последовательности, и завершает модель функция потерь CCC.

Результаты представлены в таблице 4, на валидационной и тестовой выборках результат составил 49,89% и 49,70% по показателю UAR, соответственно.

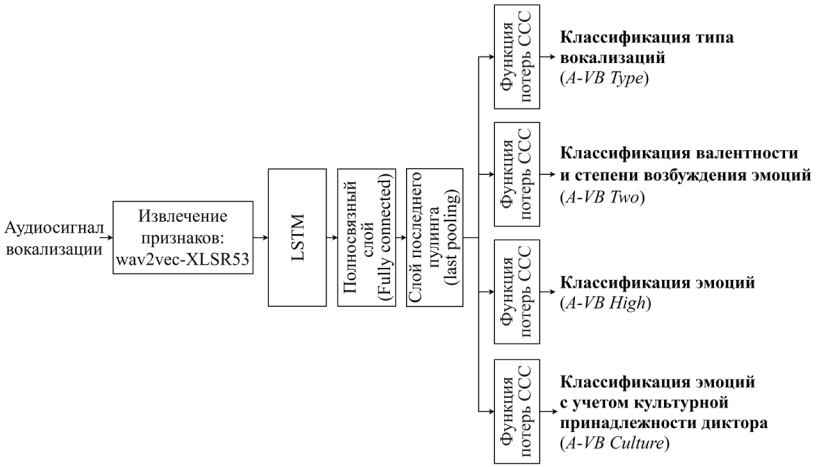

Рис. 5. Схема метода на основе wav2vec-XLSR53 в рамках соревнования A-VB'22 [40]

В рамках соревнования A-VB'22 наилучший результат показала модель data2vec-SM: 58,40% и 58,60% на валидационной и тестовой выборках, соответственно, по показателю UAR. Это единственная модель, где осуществлялась предобработка сигнала перед тем, как передать его в следующие блоки.

Однако, стоит отметить то, что во всех моделях осуществлялось параллельное прогнозирование результатов классификации по всем заданиям соревнования, и лишь метод на основе wav2vec2.0 [38], реализован в виде каскадной структуры, сначала прогнозируется результат для задания *Классификация типа вокализаций*, и затем, основываясь на результатах прогнозирования, осуществляет анализ и классификацию ЭК для каждого последующего задания. Это является правильной стратегией прогнозирования при решении нескольких задач, отличающихся сложностью и прослеживаемой иерархичностью, поскольку лишь на основе выделения общих признаков отдельно взятых ЭК, можно

предсказать более частные, например эмоцию, культурную группу диктора, и др.

Для классификации ЭК во всех методах использовались модели с уже предобученными моделями на корпусах с большим объемом данных, таких как: Multilingual LibriSpeech [41], Common Voice [42], BABEL [43], Libri–Light [45], MSP–Podcast [46], которые включают в себя записи звучащей речи как размеченной, так и неразмеченной. Модель data2vec-SM, используемая в рамках соревнования, не была дообучена после предварительного обучения на корпусе Librispeech ASR [47], с использованием инструмента для измерения скорости работы моделей Libri-Light ASR benchmark [45] для большой версии, поскольку ЭК не являются вербальной речью, а представляют собой сигнал, который состоит из вокализованных эмоций.

5. Методы автоматического анализа ЭК в потоке речи с использованием корпуса SSPNet Vocalization Corpus. Область анализа пауз хезитации и смеха, как экстралингвистических средств, в потоке речи постепенно становится более популярной, поскольку ученые осознают важность употребления подобных сигналов человеком.

Комплексные исследования ЭК в СР осуществлялись на корпусе SSPNet Vocalization Corpus, представленном в рамках соревнования *INTERSPEECH 2013 (Social Signals Sub-Challenge* [48]*)*, в котором были вручную аннотированы два типа ЭВ: смех, заполнитель или заполненные паузы хезитации - “*uhm*”, “*eh*”, “*ah*”, как сигналы удержания речевой инициативы), и «мусорные» кадры (“garbage”), которые представляют другие вокализации, такие как речь либо тишину. Корпус состоит из 2763 аудиофрагментов (11 секунд каждый): 1158 фрагментов смеха, 2988 – заполненных пауз хезитации. В рамках соревнования задача состояла в покадровой классификации двух классов событий: смеха и заполненных пауз хезитации. Точность работы метода рассчитывалась с использованием показателя площади под кривой (Area Under Curve, AUC) для каждого явления в отдельности. Поскольку распознавались два типа явлений: смех (L – laughter) и заполненные паузы-хезитации (F – fillers), расчет для двух событий (B – both) определялся невзвешенной средней оценкой площади под кривой (Unweighted Average AUC, UAAUC). Результаты исследований представлены в таблице 5.

Таблица 5. Результаты методов выявления смеха (L), заполненных пауз хезитации (F), совместно для смеха и пауз-хезитации (B)

Работа (год)	Акустические признаки	Классификатор	Результат		
			AUC		UAAUC
			L	F	B
[48] (2013)	Для MFCC 1-12, LE - Δ и Δ^2 значения. Для VP, HNR, F0, ZCR - Δ значения. μ, σ	SVM с линейным ядром	82,9	83,6	—
[50] (2014)	Для MFCC 1-12, LE - Δ и Δ^2 значения. Для VP, HNR, F0, ZCR - Δ значения. μ, σ	Deep BLSTM RNN	—	—	**94,0**
[52] (2016)	Вероятность вокализации, MFCC 1-12, HNR, F0, ZCR, логарифмическая интенсивность сигнала	DNN и вероятностное кодирование	95,3	90,4	—
[53] (2017)	MFCC	DNN+CMA-ES	94,4	88,0	91,2
	FBANK		95,0	87,7	91,3
	ComParE		**96,0**	**90,1**	93,1

5.1. Метод на основе SVM с линейным ядром [48]. В базовом подходе анализ ЭК осуществлялся с помощью метода опорных векторов с линейным ядром (linear kernel Support Vector Machines).

Набор акустических признаков для базового метода: покадрово вычислялись мел-частотные кепстральные коэффициенты (англ. Mel-Frequency Cepstral Coefficients, MFCC) 1-12 и логарифмическая энергия сигнала (logarithmic energy, LE) вместе с их производными первого (Δ) и второго порядка (Δ^2), вероятность вокализации (voicing probability, VP), отношение гармоника/шум (HNR), частота основного тона (F0), число переходов через ноль (ZCR) и их производные первого порядка. Затем для каждого низкоуровневого дескриптора (Low Level Descriptors, LLD) вычисляются среднее арифметическое (μ) и стандартное отклонение (σ) на уровне кадра и восьми соседних кадров (четыре до и четыре после). В результате вычисляется 47×3 = 141 дескриптор на кадр.

Авторы воспользовались реализацией классификатора из набора инструментов WEKA [49]. Для данной задачи использовался метод опорных векторов с линейным ядром. В качестве алгоритма обучения использовалась последовательная минимальная оптимизация (Sequential Minimal Optimization, SMO), параметр сложности, который позволяет найти компромисс между

максимизацией разделяющей полосы и минимизацией суммарной ошибки, составил $C = 10^{-3}$. Чтобы справиться с несбалансированным распределением классов, использовалась субдискретизация, после использования которой осталось только 5% «мусорных» кадров. Результат покадрового обнаружения смеха и заполненных пауз хезитации представлен в таблице 5, на тестовой выборке метод показал 82,9% и 83,6% по классам «смех» и «заполнитель», соответственно.

5.2. Метод на основе Deep BLSTM RNN [50]. В данной работе представлено использование модели двунаправленной рекуррентной нейронной сети (Reccurent Neural Network, RNN) с долгой краткосрочной памятью (Bidirectional LSTM, BiLSTM), путем наложения комбинации глубокой нейронной сети (Deep Neural Network, DNN) и BiLSTM. Схема метода представлена на рисунке 6.

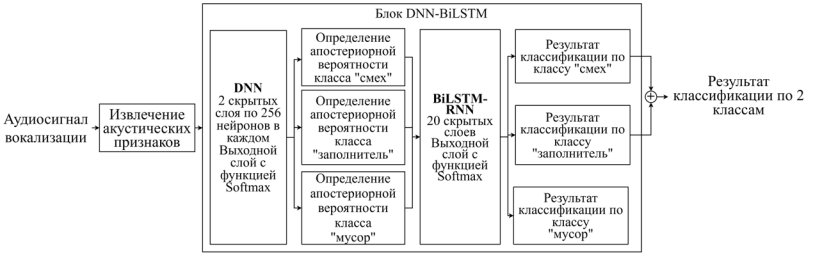

Рис. 6. Схема метода на основе DNN+BiLSTM [50]

С помощью экстрактора признаков с открытым исходным кодом openSMILE [51], каждые 10 мс при размере кадра 25 мс извлекались следующие низкоуровневые дескрипторы и функционалы: LE, коэффициенты MFCC 1-12, для которых также вычислялись Δ и Δ^2; VP, HNR, F0, ZCR и Δ для каждого признака.

Результат покадрового обнаружения смеха и заполнителей представлен в таблице 5, авторами был приведен только совместный показатель UAAUC = 94% по двум классам «смех» и «заполнитель».

5.3. Метод учета контекстных признаков и принятия решения на основе значения вероятности [52]. В работе демонстрируется трехэтапный алгоритм анализа смеха и заполненных пауз хезитации в потоке речи (рисунок 7).

Рис. 7. Схема 3-этапного метода анализа ЭК с учетом контекстных признаков и принятия решения на основе значения вероятности [52]

Этап 1. Прогнозирование вероятности события $E \in \{\text{«смех»}, \text{«заполнитель»}\}$. На этом этапе выполняются следующие действия:

а) Извлечение контекстно-независимых признаков. В данном блоке прогнозирование осуществляется с использованием предобученной DNN с двумя скрытыми слоями, на выходном слое три нейрона (в соответствии с количеством классов: «смех», «заполнитель», «мусор»), каждый с сигмоидной функцией активации прогнозирует значения вероятностей каждого из трех событий. Всего извлекались 17 признаков (таблица 5) из каждого фрейма (вектор признаков для x-го фрейма обозначен как x_n), соответствующее значение вероятности, полученное для события E в пределах фрейма, обозначается $u(n)$). Перед последующим обучением системы была осуществлена z-нормализация каждого вектора признаков для каждого файла. Средние значения и дисперсия признаков для нормализации рассчитывались с учетом всей длительности файла.

б) Учет контекстных признаков. Учет контекстных признаков осуществляется следующим образом: для прогнозирования вероятности события E в x_n-ом фрейме устанавливается длина окна $2M_x + 1$, где значение M_x – количество фреймов до и после x_n-го

фрейма. К каждому из признаков из текущего x_n-го фрейма добавляются значения Δ и Δ^2, которые извлекаются из соответствующего окна. В результате получается вектор признаков размерностью 17×3 (признак + Δ + Δ^2), затем, на наборе из 51 признака вычисляются μ и σ, и уже на данном наборе признаков реализуется дообучение DNN модели. В результате получается временной ряд временной ряд $U = (u_1(n), \dots, u_i(N))$.

Этап 2. Учет контекста в вероятностных значениях, полученных на уровне фрейма.

а) КИХ-фильтр низких частот был спроектирован для устранения выбросов во временном ряде U. Для каждого класса из E событий вычисляется сглаженное значение вероятностей $v[n]$ посредством применения сглаживающего КИХ-фильтра низких частот, как показано в формуле [52]:

$$v[n] = \sum_{m=-M}^{M} a_m u(n+m), \tag{3}$$

где a_m – это коэффициент фильтрации, применяемый к выходным фреймам на расстоянии m от текущего фрейма. Коэффициенты фильтра (a_m) определялись с помощью минимальной среднеквадратической ошибки (MMSE).

б) Вероятностное кодирование. В результате применения сглаживающего КИХ-фильтра низких частот были устранены все нелинейные зависимости. Значения вероятностей $v[n]$ поступают на вход автоэнкодера (нейронная сеть с прямой связью и сигмоидной функцией на единственном выходном нейроне). Задача данной операции, как утверждают авторы, – зафиксировать любую нелинейную контекстную зависимость, которую линейный фильтр не смог отследить. Детальное описание модели автоэнкодера в статье не приведено.

Этап 3. Маскирование высоко- и мало-вероятностных событий на основе правил, основанных на эвристике:

1) маскирование нулями: если существует непрерывное существование значений вероятности $w_E(n)$ ниже порогового значения T_0, по крайней мере, для заданного числа фреймов K_0, все такие вероятности маскируются нулем.

2) маскирование единицами: если значения вероятности смежно превышают пороговое значение T_1, по крайней мере, для числа K_1 фреймов, все такие вероятности маскируются единицей.

Подбор пороговых значений T_0 и T_1, и значения фреймов K_0 и K_1 осуществлялся на валидационном наборе данных. В [52] значения $T_0 = 0{,}02$ и $T_1 = 0{,}98$.

Результат пофреймового обнаружения смеха и пауз хезитации представлен в таблице 5, на тестовой выборке метод показал AUC = 95,3% и 90,4% по классам «смех» и «заполнитель», соответственно.

5.4. Метод на основе DNN и CMA-ES [53]. В данной статье приведен пример использования DNN и фильтра, сглаживающего временные ряды, оптимизация весов которого была проведена на основе метода оптимизации CMA-ES (Covariance Matrix Adaptation Evolution Strategy – эволюционная стратегия адаптации ковариационной матрицы). Метод оптимизации весов фильтра с помощью CMA-ES подробно описан в [54] и имеет сходство с методом главных компонент.

Авторы выделили три группы признаков, описанные ниже, на основе которых проводилось покадровый анализ фрагментов СР:

1) Стандартный: 13 коэффициентов MFCC, их первая и вторая производные.

2) FBANK: 20 значений гребенки мел-частотных фильтров, логарифмическая энергия, значения Δ и Δ^2.

3) ComParE – *INTERSPEECH 2013 Computational Paralinguistic Challenge*: 13 коэффициентов MFCC, значения Δ и Δ^2; VP, HNR, F0, ZCR и Δ для каждого признака. В общей сложности 47 признаков, которые извлекаются с помощью набора программных средств openSMILE [51].

Пофреймовые значения вероятностей были получены с помощью DNN, изначально обученной для получения апостериорной оценки фонемы на уровне фрейма, методом скользящего окна с учетом соседних векторов признаков на уровне фреймов. Подробное описание обучения DNN в статье не приведено.

Полученные значения вероятностей на каждом фрейме агрегируются в локальном окружении из нескольких векторов, число которых было определено на предыдущем этапе. Для фильтра с шириной $2N + 1$ значения весов были определены следующим образом: $(w_{-N}, w_{-N+1}, \dots, w_N) \geq 0$ и $\sum_{i=-N}^{N} w_i = 1$. Далее для каждого j-го фрейма со значением правдоподобия вычислялось значение агрегирования по формуле [53]:

$$a_j^{'} = \sum_{i=-N}^{N} w_i a_{j+i}, \quad (4)$$

где w_i – значения весов, которые подбираются с помощью определенного типа фильтра.

В общей сложности были протестированы 4 типа фильтра: константный, треугольный, фильтр на основе DNN, и фильтр на основе CMA-ES. Лучший результат показал фильтр, значения весов которого были подобраны с помощью метода CMA-ES.

Результат покадрового обнаружении смеха и заполнителей представлен в таблице 5, на тестовой выборке метод показал 95,0% и 90,1% по классам «смех» и «заполнитель», соответственно. Результат для совместных событий составил UAAUC =93,1%.

Наилучший результат UAAUC = 94,0% показала модель на основе комбинации DNN и BiLSTM, однако, в статье не приведены отдельные результаты по классам «смех» и «заполнители», лучшие отдельные результаты показала модель на основе DNN и оптимизатора CMA-ES – 96,0% по классу «смех» и 90,1% по классу «заполнитель».

На основе проведенного обзора литературы можно сделать вывод, что повышение качества анализа ЭК в СР возможно при условии использования фильтров сглаживания временных рядов после выделения высокоуровневых признаков с использованием нейронной сети. Также, стоит отметить, что при анализе ЭК в СР необходимо принимать во внимание признаки не только анализируемого кадра, но и нескольких соседних (т.е. осуществлять анализ контекста), для повышения точности используемого метода. Стоит отметить, что ни в одном из вышеперечисленных подходов не использовались методы предобработки речевого сигнала для последующего извлечения признаков. Материал корпуса SSPNet Vocalization Corpus был записан на мобильное устройство Nokia N900, что могло повлиять на качество записанного речевого материала.

6. Заключение. В данной статье представлен аналитический обзор научных исследований, посвященных анализу ЭК в СР, так и ЭК, записанных отдельно. Рассмотрены лингвистические основы ЭК, приведена оригинальная систематизация. Представлен сравнительно-сопоставительный обзор корпусов со СР и наборы данных с ЭС, записанными отдельно от речи.

На основе проведенного аналитического обзора существующих подходов к анализу ЭК в СР, можно сформулировать следующие

предложения для повышения точности распознавания подобных явлений в СР:

а. При аннотации ЭК в корпусах СР важно размечать фрагменты пересечения речи и ЭК. ЭК могут проявляться не только как отдельные от речи события, но и пересекаться с речью и другими средствами. Есть предположение, что в случае, отсутствия пересечения ЭК с речью точность анализа ЭК в потоке СР будет проходить успешнее, нежели, если речь смешивается с ЭК. И, как следствие, репрезентативное представление пересечений ЭК с речью в корпусах СР позволит повысить точность анализа подобных событий в потоке СР.

б. При исследовании ЭК, которые включают в себя два рефлекторных процесса (вдох и выдох), такие как смех, плач, крик и др., целесообразно учитывать акустические характеристики как вдоха, так и выдоха, поскольку представляют одно цельное событие (смех/плач/крик), а не отдельные друг от друга явления.

в. Первично исследование ЭК в рамках одной культуры, поскольку авторы предполагают межкультурную вариативность в реализациях ЭК. Межкультурное исследование ЭК вторично, поскольку возможны схожие реализации отдельных ЭК в рамках одной языковой группы.

г. В условиях зашумленной обстановки качество выявления и анализа ЭК значительно снижается, вследствие чего необходимо производить предварительную очистку или подавление, отличных от речи шумов, присутствующих в сигнале.

д. Для повышения точности распознавания ЭК в СР необходимо рассматривать более широкий спектр корпусов и баз данных. Для обучения системы необходимо использовать не только записи СР, но и отдельно записанные ЭК, также возможно использование записей людей с дефектами речи и моторно-речевыми нарушениями, тем самым, выборка реализаций ЭК будет более вариативна и станет возможным учет различных особенностей воспроизводства тех или иных ЭК.

е. Для более точного выявления ЭК необходимо либо увеличивать ширину фрейма, либо при анализе текущего фрейма учитывать контекст (3-4 соседних фрейма), связано это с тем, что скорость смены фонем не соответствует скорости возникновения и продолжительности ЭК в потоке речи, следовательно, для фиксирования ЭК либо повторяемости акустических признаков, похожих на ЭК в сигнале методу необходимо захватывать «больший обзор» сигнала для его анализа.

В дальнейшей работе планируется проведение ряда исследований, посвященных изучению акустических характеристик различных ЭК, исследование статистических зависимостей, и анализ ЭК в потоке СР.

1) Планируется провести исследования на наборах данных с ЭК, записанными отдельно от речи: «Multiclass cough sound dataset» [7], «ASVP-ESD» [31], FSD50K (Human Sounds) [34], «VocalSound» [35], для того чтобы подобрать оптимальные модели искусственных нейронных сетей (ИНС), которые позволят повысить точность распознавания ЭК в формате отдельных от речи аудиофрагментов. Планируется осуществить сопоставление и тестирование различных наборов признаков (GeMAPS, eGeMAPS, BoAW, wav2vec, wav2vec2, и др.) на качество распознавания ЭК.

Также, для повышения точности распознавания отдельно записанных ЭК будет необходимо провесит аугментацию данных, путем извлечения различных реализаций ЭК из корпусов СР, также возможна запись уникального корпуса СР и отдельно записанных ЭК.

2) Анализ реализаций ЭК в рамках спонтанной русской речи. Исследование будет проводиться на корпусах подобных «CAT» [17], «Корпуса звучащей речи» [19], «CORUSS» [20]. Предполагается исследовать методы предобработки речевого сигнала со СР для последующего выделения векторов признаков из каждого фрейма. В рамках второго этапа планируется изучить влияние ширины фрейма на качество распознавания ЭК в СР.

3) Кросс-культурное исследование ЭК. Исследование возможности применения методов на основе ИНС, разработанных для анализа ЭК в рамках одной культуры (русский язык), для распознавания ЭК в другой (английский язык). В рамках исследования ЭК у англоговорящих дикторов возможно использование следующих корпусов со СР: Buckeye [22], LUCID [23], Wildcat [25], SpiCE [26], ICSI meeting corpus [28].

4) Оформление выводов по произведенной работе и разработка автоматической системы, которая будет осуществлять анализ ЭК различного типа в СР.

Литература

1. Верходанова В.О., Шапранов В.В., Кипяткова И.С., Карпов А.А. Автоматическое определение вокализованных хезитаций в русской речи // Вопросы языкознания. 2018. № 6. С. 104–118.
2. Ataollahi F., Suarez M.T. Laughter Classification Using 3D Convolutional Neural Networks // Proceedings of the 3rd International Conference on Advances in Artificial Intelligence (ICAAI '19). 2019. pp. 47–51.

3. Судьенкова А.В. Обзор методов извлечения акустических признаков речи в задаче распознавания диктора // Сборник научных трудов НГТУ. 2019. № 3–4. С. 139–164.
4. Hsu J.-H., Su M.-H., Wu C.-H., Chen Y.-H. Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. vol. 29. pp. 1675–1686.
5. Dumpala S.H., Alluri K.N.R.K.R. An Algorithm for Detection of Breath Sounds in Spontaneous Speech with Application to Speaker Recognition. Speech and Computer: 19th International Conference (SPECOM). 2017. pp. 98–108.
6. Huang K.-Y., Wu C.-H., Hong Q.-B., Su M.-H., Chen Y.-H. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. pp. 5866–5870.
7. Kuluozturk M., Kobat M.A., Barua P.D., Dogan S., Tuncer T., Tan R.S., Ciaccio E.J., Acharya U.R. DKPNet41: Directed knight pattern network-based cough sound classification model for automatic disease diagnosis // Medical engineering and physics. 2022. vol. 110. no. 103870.
8. Lahmiri S., Tadj C., Gargour C., Bekiros S. Deep learning systems for automatic diagnosis of infant cry signals // Chaos, Solitons & Fractals. 2022. vol. 154. no. 111700.
9. Matikolaie F.S., Tadj C. Machine Learning-Based Cry Diagnostic System for Identifying Septic Newborns // Journal of Voice. 2022. DOI: 10.1016/j.jvoice.2021.12.021.
10. Matsuda T., Arimoto Y. Detection of laughter and screaming using the attention and ctc models // Proceedings of INTERSPEECH 2023. pp. 1025–1029. DOI: 10.21437/Interspeech.2023-1412.
11. Ortega D., Meyer S., Schweitzer A., Vu N.T. Modeling Speaker-Listener Interaction for Backchannel Prediction // 13th International Workshop on Spoken Dialogue Systems Technology. 2023. pp. 1–16.
12. Lea C., Huang Z., Jain D., Tooley L., Liaghat Z., Thelapurath S., Findlater L., Bigham J.P. Nonverbal Sound Detection for Disordered Speech // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 7397–7401.
13. Crystal D. Prosodic Systems and Intonation in English // Cambridge University Press, 1969. 390 p.
14. Simon-Thomas E., Sauter D., Sinicropi-Yao L., Abramson A., Keltner D. Vocal Bursts Communicate Discrete Emotions: Evidence for New Displays. Nature Proceedings. 2007. DOI: 10.1038/npre.2007.1356.1.
15. Trouvain J., Truong K.P. Comparing non-verbal vocalisations in conversational speech corpora. Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3'2012). 2012. pp. 36–39.
16. Савельева Н.А., Пальчик А.Б., Калашникова Т.П. Особенности довербальной вокализации у плодов и младенцев // Специальное образование. 2022. № 2(66). С. 246–259.
17. Богданова-Бегларян Н.В., Блинова О.В., Зайдес К.Д., Шерстинова Т.Ю. Корпус «Сбалансированная аннотированная текстотека» (CAT): изучение специфики русской монологической речи // Труды института русского языка им. В.В. Виноградова. 2019. № 21. С. 110–126.
18. Богданова-Бегларян Н.В., Шерстинова Т.Ю., Блинова О.В., Мартыненко Г.Я. Корпус «Один речевой день» в исследованиях социолингвистической

вариативности русской разговорной речи // Анализ разговорной русской речи (АРЗ – 2017): труды седьмого междисциплинарного семинара Санкт-Петербург. 2017. С. 14–20.
19. Кибрик А.А., Подлесская В.И. Коррекция в устной русской монологической речи по данным корпусного исследования // Русский язык в научном освещении. 2006. № 2. С. 7–55.
20. Kachkovskaia T., Kocharov D., Skrelin P., Volskaya N. CoRuSS – a New Prosodically Annotated Corpus of Russian Spontaneous Speech // Proceedings of the tenth international conference on language resources and evaluation. Portoroz, Slovenia. 2016. pp. 1949–1954.
21. Кибрик А.А. Русский мультиканальный дискурс. Часть II. Разработка корпуса и направления исследований // Психологический журнал. 2018. № 39(2). С. 79–90.
22. Pitt M.A., Johnson K., Hume E., Kiesling S., Raymond W. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability // Speech Communication. 2005. vol. 45(1). no. 1. pp. 89–95.
23. Baker R., Hazan V. LUCID: a corpus of spontaneous and read clear speech in British English // Proceedings of DiSS-LPSS Joint Workshop. 2010. pp. 3–6.
24. Polychroniou A., Salamin H., Vinciarelli A. The SSPNet-Mobile Corpus: Social Signal Processing Over Mobile Phones // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 1492–1498.
25. Van Engen K.J., Baese-Berk M., Baker R.E., Choi A., Kim M., Bradlow A.R. The Wildcat Corpus of native- and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles // Language and speech. 2010. vol. 53(4). pp. 510–540.
26. Johnson K.A., Babel M., Fong I., Yiu N. SpiCE: A New Open-Access Corpus of Conversational Bilingual Speech in Cantonese and English // Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association (ELRA). 2020. pp. 4089–4095.
27. Baese-Berk M.M., Morrill T.H. Speaking rate consistency in native and non-native speakers of English // The Journal of the Acoustical Society of America. 2015. vol. 138(3). pp. 223–228.
28. Janin A., Baron D., Edwards J., Ellis D., Gelbart D., Morgan N., Wooters C. The ICSI Meeting Corpus // IEEE International Conference on Acoustics, Speech, and Signal Processing. 2003. vol. 1. DOI: 10.1109/icassp.2003.1198793.
29. Chou H.C., Lin W.C., Chang L.C., Li C.C., Ma H.P., Lee C.C. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus // Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII'2017). 2017. pp. 292–298.
30. Gosy M. BEA – a multifunctional Hungarian spoken language data base // The Phonetician. 2012. vol. 105. pp. 50–61.
31. Landry Dejoli T.T., He Q., Yan H., Li Y. ASVP-ESD: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances // Global Scientific Journals. 2020. vol. 8(5). pp. 1793–1798.
32. Baird A., Tzirakis P., Brooks J.A., Gregory C.B., Schuller B., Batliner A., Keltner D., Cowen A. The ACII 2022 Affective Vocal Bursts Workshop & Competition: Understanding a critically understudied modality of emotional expression // 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2022.
33. Petridis S., Martinez B., Pantic M. The MAHNOB Laughter database // Image and Vision Computing. 2013. vol. 31(2). pp. 186–202.

34. Fonseca E., Favory X., Pons J., Font F., Serra X. FSD50K: An Open Dataset of Human-Labeled Sound Events // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022. vol. 30. pp. 829–852.
35. Gong Y., Yu J., Glass J. Vocalsound: A Dataset for Improving Human Vocal Sounds Recognition // International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 151–155.
36. Kantharaju R.B., Ringeval F., Besacier L. Automatic recognition of affective laughter in spontaneous dyadic interactions from audiovisual signals // Proceedings of the ACM 20th International Conference on Multimodal Interaction (ICMI'18). 2018. pp. 220–228.
37. Hallmen T., Mertes S., Schiller D., André E. An efficient multitask learning architecture for affective vocal burst analysis // arXiv preprint arXiv: abs/2209.13914. 2022.
38. Karas V., Triantafyllopoulos A., Song M., Schuller B.W. Self-Supervised Attention Networks and Uncertainty Loss Weighting for Multi-Task Emotion Recognition on Vocal Bursts // The 2022 ACII Affective Vocal Burst Workshop & Challenge (A-VB). 2022. vol. 45(1). pp. 89–95.
39. Liu S., Johns E., Davison A.J. End-to-end multi-task learning with attention // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. pp. 1871–1880.
40. Nguyen D.-K., Pant S., Ho N.-H., Lee G.-S., Kim S.-H., Yang H.-J. Fine-tuning Wav2vec for Vocal-burst Emotion Recognition // The 2022 ACII Affective Vocal Burst Workshop & Challenge (A-VB). 2022. vol. 45(1). pp. 89–95.
41. Pratap V., Xu Q., Sriram A., Synnaeve G., Collobert R. MLS: a large-scale multilingual dataset for speech research // Proceedings of INTERSPEECH. 2020. pp. 2757–2761.
42. Ardila R., Branson M., Davis K., Henretty M., Kohler M., Meyer J., Morais R., Saunders L., Tyers F.M., Weber G. Common voice: a massively-multilingual speech corpus // Proceedings of the 12th Conference on Language Resources and Evaluation (LREC'2020). 2020. pp. 4218–4222.
43. Gales M.J.F., Knill K., Ragni A., Rath S.P. Speech recognition and keyword spotting for low-resource languages: babel project research at cued // Proceedings 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU'2014). 2012. pp. 16–23.
44. Vaessen N., Van Leeuwen D.A. Fine-tuning wav2vec2 for speaker recognition // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 7967–7971.
45. Kahn J., Riviere M, Zheng W., Kharitonov E., Xu Q., Mazare P-E., Karaday J., Liptchinsky V., Collobert R., Fuegen C., et al. Libri-light: A benchmark for asr with limited or no supervision // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. pp. 7669–7673.
46. Lotfian R., Busso C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings // IEEE Transactions on Affective Computing. 2019. vol. 10. no. 4. pp. 471–483.
47. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: an ASR corpus based on public domain audio books // IEEE international conference on acoustics, speech and signal processing (ICASSP). 2015. pp. 5206–5210.
48. Schuller B., Steidl S., Batliner A., Vinciarelli A., Scherer K., Ringeval F., Chetouani M., Weninger F., Eyben F., Marchi E., Mortillaro M., Salamin H., Polychroniou A., Valente F., Kim S. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism // Proceedings of the 14th Annual

Conference of the International Speech Communication Association. 2013. pp. 148–152.
49. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. The WEKA data mining software: An update // ACM SIGKDD Explorations Newsletter. 2009. vol. 11. no. 1. pp. 10–18.
50. Brueckner R., Schuller B. Social signal classification using deep BLSTM recurrent neural networks // International conference on acoustics, speech and signal processing (ICASSP). 2014. pp. 4823–4827.
51. Eyben F., Wollmer M., Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor // Proceedings 18th ACM International Conference Multimedia. 2010. pp. 1459–1462.
52. Gupta R., Audhkhasi K., Lee S., Narayanan S. Detecting paralinguistic events in audio stream using context in features and probabilistic decisions // Computer Speech & Language. 2016. vol. 36. pp. 72–92.
53. Gosztolya G. Optimized Time Series Filters for Detecting Laughter and Filler Events // INTERSPEECH. 2017. pp. 2376–2380.
54. Hansenand N., Ostermeier A. Completely derandomized selfadaptation in evolution strategies // Evolutionary Computation. 2001. vol. 9. no. 2. pp. 159–195.

Поволоцкая Анастасия Андреевна — младший научный сотрудник, аспирант, лаборатория речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: искусственный интеллект, машинное обучение, нейронные сети, автоматическое распознавание речи, компьютерная лингвистика. Число научных публикаций — 3. anastasiia.povolotskaia@gmail.com; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Карпов Алексей Анатольевич — д-р техн. наук, профессор, руководитель лаборатории, лаборатория речевых и многомодальных интерфейсов, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: речевые технологии, автоматическое распознавание речи, обработка аудиовизуальной речи, многомодальные человеко-машинные интерфейсы, компьютерная паралингвистика и другие. Число научных публикаций — 350+. karpov@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Поддержка исследований. Данное исследование выполнено в рамках Ведущей научной школы РФ (Грант № НШ-17.2022.1.6), а также частично в рамках бюджетной темы СПб ФИЦ РАН (№ FFZF-2022-0005).

DOI 10.15622/ia.23.1.1

A. POVOLOTSKAIA, A. KARPOV

ANALYTICAL REVIEW OF METHODS FOR AUTOMATIC ANALYSIS OF EXTRA-LINGUISTIC UNITS IN SPONTANEOUS SPEECH

Povolotskaia A., Karpov A. **Analytical Review of Methods for Automatic Analysis of Extra-Linguistic Units in Spontaneous Speech.**

Abstract. The accuracy of automatic spontaneous speech recognition systems is far from that of trained speech recognition systems. This is due to the fact that spontaneous speech is not as smooth and failure-free as spontaneous speech. Spontaneous speech varies from speaker to speaker: the quality of phonemes' pronunciation, the presence of pauses, speech disruptions and extralinguistic items (laughing, coughing, sneezing, and chuckling when expressing emotions of irritation, etc.) interrupt the fluency of verbal speech. However, it is worth noting that extralinguistic items very often carry important paralinguistic information, so it is crucial for automatic spontaneous speech recognition systems not only to identify such phenomena and distinguish them from the verbal components of speech but also to classify them. This review presents an analysis of works on the topic of automatic detection and analysis of extralinguistic items in spontaneous speech. Both individual methods and approaches to the recognition of extralinguistic items in a speech stream, and works related to the multiclass classification of isolatedly recorded extralinguistic units are considered and described. The most popular methods of extralinguistic units' analysis are neural networks, such as deep neural networks and networks based on transformer models. The basic concepts related to the term extralinguistic items are given, the original systematization of extralinguistic items in the Russian language is proposed, the corpus and databases of audio spoken speech both in Russian and in other languages are described, the data sets of extralinguistic items recorded isolatedly are also given. The accuracy of extralinguistic items recognition increases with the following conditions of work with the speech signal: pre-processing of audio signals of items has shown an increase in the accuracy of separately recorded extralinguistic items classification; consideration of context (analysis of several frames of speech signal) and use of filters for smoothing the time series after extraction of feature vectors showed an increase in accuracy in frame-by-frame analysis of the speech signal with spontaneous speech.

Keywords: automatic speech recognition, speech technology, machine learning, linguistics, extralinguistic units, spontaneous speech, automatic extralinguistic units recognition.

References

1. Verkhodanova V., Karpov A., Kipyatkova I., Shapranov V. [Automatic detection of vocalized hesitations in Russian speech]. Voprosy Jazykoznanija – Questions of linguistics. 2018. no. 6. pp. 104–118. (In Russ.).
2. Ataollahi F., Suarez M.T. Laughter Classification Using 3D Convolutional Neural Networks. Proceedings of the 3rd International Conference on Advances in Artificial Intelligence (ICAAI '19). 2019. pp. 47–51.
3. Sudjenkova A.V. [Overview of methods for extracting acoustic speech features in speaker recognition]. Sbornik nauchnyh trudov NGTU – Transaction of scientific papers of the Novosibirsk state technical university. 2019. no. 3–4. pp. 139–164. (In Russ.).

4. Hsu J.-H., Su M.-H., Wu C.-H., Chen Y.-H. Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. vol. 29. pp. 1675–1686.
5. Dumpala S.H., Alluri K.N.R.K.R. An Algorithm for Detection of Breath Sounds in Spontaneous Speech with Application to Speaker Recognition. Speech and Computer: 19th International Conference (SPECOM). 2017. pp. 98–108.
6. Huang K.-Y., Wu C.-H., Hong Q.-B., Su M.-H., Chen Y.-H. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. pp. 5866–5870.
7. Kuluozturk M., Kobat M.A., Barua P.D., Dogan S., Tuncer T., Tan R.S., Ciaccio E.J., Acharya U.R. DKPNet41: Directed knight pattern network-based cough sound classification model for automatic disease diagnosis. Medical engineering and physics. 2022. vol. 110. no. 103870.
8. Lahmiri S., Tadj C., Gargour C., Bekiros S. Deep learning systems for automatic diagnosis of infant cry signals. Chaos, Solitons & Fractals. 2022. vol. 154. no. 111700.
9. Matikolaie F.S., Tadj C. Machine Learning-Based Cry Diagnostic System for Identifying Septic Newborns. Journal of Voice. 2022. DOI: 10.1016/j.jvoice.2021.12.021.
10. Matsuda T., Arimoto Y. Detection of laughter and screaming using the attention and ctc models. Proceedings of INTERSPEECH 2023. pp. 1025–1029. DOI: 10.21437/Interspeech.2023-1412.
11. Ortega D., Meyer S., Schweitzer A., Vu N.T. Modeling Speaker-Listener Interaction for Backchannel Prediction. 13th International Workshop on Spoken Dialogue Systems Technology. 2023. pp. 1–16.
12. Lea C., Huang Z., Jain D., Tooley L., Liaghat Z., Thelapurath S., Findlater L., Bigham J.P. Nonverbal Sound Detection for Disordered Speech. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 7397–7401.
13. Crystal D. Prosodic Systems and Intonation in English. Cambridge University Press, 1969. 390 p.
14. Simon-Thomas E., Sauter D., Sinicropi-Yao L., Abramson A., Keltner D. Vocal Bursts Communicate Discrete Emotions: Evidence for New Displays. Nature Proceedings. 2007. DOI: 10.1038/npre.2007.1356.1.
15. Trouvain J., Truong K.P. Comparing non-verbal vocalisations in conversational speech corpora. Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3'2012). 2012. pp. 36–39.
16. Savel'eva N.A., Pal'chik A.B., Kalashnikova T.P. [Specific features of preverbal vocalizations in fetuses and infants]. Special'noe obrazovanie – Special Education. 2022. no. 2(66). pp. 246–259. (In Russ.).
17. Bogdanova-Beglarjan N.V., Blinova O.V., Zajdes K.D., Sherstinova T.Ju. [Corpus "Balanced annotated text collection (textoteca)" (SAT): studying the specificity of russian monological speech]. Trudy instituta russkogo jazyka im. V.V. Vinogradova – Proceedings of the V.V. Vinogradov Institute of Russian Language. Vinogradov. 2019. no. 21. pp. 110–126. (In Russ.).
18. Bogdanova-Beglarjan N.V., Sherstinova T.Ju., Blinova O.V., Martynenko G.Ja. [Corpus «One Speech Day» in studies of sociolinguistic variation in Russian colloquial speech] Analiz razgovornoj russkoj rechi (AR3 – 2017): Trudy sed'mogo

mezhdisciplinarnogo seminara [Analysis of the Russian colloquial speech: Collected papers]. St Petersburg: St Petersburg State University, 2017. pp. 14–20. (In Russ.).
19. Kibrik A.A., Podlesskaja V.I. Correction in Russian spoken monologues: a corpus study. Russkii yazyk v nauchnom osveshchenii – Russian Language and Linguistic Theory. 2006. no. 2. pp. 7–55. (In Russ.).
20. Kachkovskaia T., Kocharov D., Skrelin P., Volskaya N. CoRuSS – a New Prosodically Annotated Corpus of Russian Spontaneous Speech. Proceedings of the tenth international conference on language resources and evaluation. Portoroz, Slovenia. 2016. pp. 1949–1954.
21. Kibrik A.A. [Russian Multichannel Discourse. Part II. Development of the corpus and directions of research]. Psihologicheskij zhurnal – Journal of Psychology. 2018. no. 39(2). pp. 79–90. (In Russ.).
22. Pitt M.A., Johnson K., Hume E., Kiesling S., Raymond W. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication. 2005. vol. 45(1). no. 1. pp. 89–95.
23. Baker R., Hazan V. LUCID: a corpus of spontaneous and read clear speech in British English. Proceedings of DiSS-LPSS Joint Workshop. 2010. pp. 3–6.
24. Polychroniou A., Salamin H., Vinciarelli A. The SSPNet-Mobile Corpus: Social Signal Processing Over Mobile Phones. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. pp. 1492–1498.
25. Van Engen K.J., Baese-Berk M., Baker R.E., Choi A., Kim M., Bradlow A.R. The Wildcat Corpus of native- and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles. Language and speech. 2010. vol. 53(4). pp. 510–540.
26. Johnson K.A., Babel M., Fong I., Yiu N. SpiCE: A New Open-Access Corpus of Conversational Bilingual Speech in Cantonese and English. Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association (ELRA). 2020. pp. 4089–4095.
27. Baese-Berk M.M., Morrill T.H. Speaking rate consistency in native and non-native speakers of English. The Journal of the Acoustical Society of America. 2015. vol. 138(3). pp. 223–228.
28. Janin A., Baron D., Edwards J., Ellis D., Gelbart D., Morgan N., Wooters C. The ICSI Meeting Corpus. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2003. vol. 1. DOI: 10.1109/icassp.2003.1198793.
29. Chou H.C., Lin W.C., Chang L.C., Li C.C., Ma H.P., Lee C.C. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII'2017). 2017. pp. 292–298.
30. Gosy M. BEA – a multifunctional Hungarian spoken language data base. The Phonetician. 2012. vol. 105. pp. 50–61.
31. Landry Dejoli T.T., He Q., Yan H., Li Y. ASVP-ESD: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances. Global Scientific Journals. 2020. vol. 8(5). pp. 1793–1798.
32. Baird A., Tzirakis P., Brooks J.A., Gregory C.B., Schuller B., Batliner A., Keltner D., Cowen A. The ACII 2022 Affective Vocal Bursts Workshop & Competition: Understanding a critically understudied modality of emotional expression. 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2022.
33. Petridis S., Martinez B., Pantic M. The MAHNOB Laughter database. Image and Vision Computing. 2013. vol. 31(2). pp. 186–202.

34. Fonseca E., Favory X., Pons J., Font F., Serra X. FSD50K: An Open Dataset of Human-Labeled Sound Events. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022. vol. 30. pp. 829–852.
35. Gong Y., Yu J., Glass J. Vocalsound: A Dataset for Improving Human Vocal Sounds Recognition. International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 151–155.
36. Kantharaju R.B., Ringeval F., Besacier L. Automatic recognition of affective laughter in spontaneous dyadic interactions from audiovisual signals. Proceedings of the ACM 20th International Conference on Multimodal Interaction (ICMI'18). 2018. pp. 220–228.
37. Hallmen T., Mertes S., Schiller D., André E. An efficient multitask learning architecture for affective vocal burst analysis. arXiv preprint arXiv: abs/2209.13914. 2022.
38. Karas V., Triantafyllopoulos A., Song M., Schuller B.W. Self-Supervised Attention Networks and Uncertainty Loss Weighting for Multi-Task Emotion Recognition on Vocal Bursts. The 2022 ACII Affective Vocal Burst Workshop & Challenge (A-VB). 2022. vol. 45(1). pp. 89–95.
39. Liu S., Johns E., Davison A.J. End-to-end multi-task learning with attention. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. pp. 1871–1880.
40. Nguyen D.-K., Pant S., Ho N.-H., Lee G.-S., Kim S.-H., Yang H.-J. Fine-tuning Wav2vec for Vocal-burst Emotion Recognition. The 2022 ACII Affective Vocal Burst Workshop & Challenge (A-VB). 2022. vol. 45(1). pp. 89–95.
41. Pratap V., Xu Q., Sriram A., Synnaeve G., Collobert R. MLS: a large-scale multilingual dataset for speech research. Proceedings of INTERSPEECH. 2020. pp. 2757–2761.
42. Ardila R., Branson M., Davis K., Henretty M., Kohler M., Meyer J., Morais R., Saunders L., Tyers F.M., Weber G. Common voice: a massively-multilingual speech corpus. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC'2020). 2020. pp. 4218–4222.
43. Gales M.J.F., Knill K., Ragni A., Rath S.P. Speech recognition and keyword spotting for low-resource languages: babel project research at cued. Proceedings 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU'2014). 2012. pp. 16–23.
44. Vaessen N., Van Leeuwen D.A. Fine-tuning wav2vec2 for speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 7967–7971.
45. Kahn J., Riviere M, Zheng W., Kharitonov E., Xu Q., Mazare P-E., Karaday J., Liptchinsky V., Collobert R., Fuegen C., et al. Libri-light: A benchmark for asr with limited or no supervision. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. pp. 7669–7673.
46. Lotfian R., Busso C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEE Transactions on Affective Computing. 2019. vol. 10. no. 4. pp. 471–483.
47. Panayotov V., Chen G., Povey D., Khudanpur S. LibriSpeech: an ASR corpus based on public domain audio books. IEEE international conference on acoustics, speech and signal processing (ICASSP). 2015. pp. 5206–5210.
48. Schuller B., Steidl S., Batliner A., Vinciarelli A., Scherer K., Ringeval F., Chetouani M., Weninger F., Eyben F., Marchi E., Mortillaro M., Salamin H., Polychroniou A., Valente F., Kim S. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. Proceedings of the 14th Annual

Conference of the International Speech Communication Association. 2013. pp. 148–152.
49. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter. 2009. vol. 11. no. 1. pp. 10–18.
50. Brueckner R., Schuller B. Social signal classification using deep BLSTM recurrent neural networks. International conference on acoustics, speech and signal processing (ICASSP). 2014. pp. 4823–4827.
51. Eyben F., Wollmer M., Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor. Proceedings 18th ACM International Conference Multimedia. 2010. pp. 1459–1462.
52. Gupta R., Audhkhasi K., Lee S., Narayanan S. Detecting paralinguistic events in audio stream using context in features and probabilistic decisions. Computer Speech & Language. 2016. vol. 36. pp. 72–92.
53. Gosztolya G. Optimized Time Series Filters for Detecting Laughter and Filler Events. INTERSPEECH. 2017. pp. 2376–2380.
54. Hansenand N., Ostermeier A. Completely derandomized selfadaptation in evolution strategies. Evolutionary Computation. 2001. vol. 9. no. 2. pp. 159–195.

Povolotskaia Anastasiia — Junior researcher, postgraduate student, Speech and multimodal interfaces laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: artificial intelligence, machine learning, neural networks, automatic speech recognition, computer linguistics. The number of publications — 3. anastasiia.povolotskaia@gmail.com; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Karpov Alexey — Ph.D., Dr.Sci., Professor, Head of the laboratory, Speech and multimodal interfaces laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: speech technology, automatic speech recognition, audio-visual speech processing, multimodal human-computer interfaces, and computational paralinguistics. The number of publications — 350+. karpov@iias.spb.su; 39, 14-th Line V.O., 199178, Saint-Petersburg, Russia; office phone: +7(812)328-0421.

Acknowledgements. This research was partially supported by the Leading scientific school of Russia (Grant no. NSH-17.2022.1.6), as well as partially by the state research of SPC RAS (Topic no. FFZF-2022-0005).