**УДК 811.161.1’33’342+159.942.33**

**У. Е. Кочеткова, П. А. Скрелин, П. П. Щербаков**
г. Санкт-Петербург, Россия,
Санкт-Петербургский государственный университет

# ОПРЕДЕЛЕНИЕ ЭМОЦИЙ В АУДИОВИЗУАЛЬНОМ СИГНАЛЕ

Цель исследования состояла в разработке методов автоматического определения эмоций в аудиовизуальном сигнале на материале русского языка. В связи с неточностью определения эмоциональной окраски существующими инструментами автоматической обработки видеосигнала, была обучена модель распознавания мимических движений на материале мультимедийного корпуса русской иронической речи, содержащего фонетическую аннотацию, а также оценку наличия или отсутствия иронии носителями языка. В работе сравниваются данные акустического анализа и анализа видеоряда в иронических и неиронических высказываниях.

*Ключевые слова*: акустический анализ; математические модели; мимика; речевой корпус; ирония; просодические характеристики; фонетические единицы.

**U. Y. Kochetkova, P. A. Skrelin, P. P. Scherbakov**
St. Petersburg, Russia, St. Petersburg State University

# DEFINING EMOTIONS IN AN AUDIO-VISUAL SIGNAL

The purpose of the study was to develop the methods for the automatic detection of emotions in an audio-visual signal based on the material of the Russian language. Due to the insufficient accuracy of determining emotions by facial expressions through automatic detection, a model was trained based on the material of the multimedia corpus of Russian ironic speech, which contains phonetic annotation and evaluation of ironic or non-ironic meaning by native speakers. The paper compares the data of acoustic analysis and video sequence analysis in ironic and non-ironic statements.

*Keywords*: acoustic analysis; mathematical models; facial expressions; speech corpus; irony; prosodic characteristics; phonetic units.

В настоящее время наблюдается стремительное развитие систем искусственного интеллекта и их внедрение в различные сферы жизни, при этом всё активнее используются аудиовизуальные интерфейсы, предполагающие сочетание синтеза звукового сигнала с синтезом изображения. Однако до сих пор данные системы являются несовершенными: по сравнению с синтезом на уровне отдельных фонем, интонационный синтез часто бывает неестественным, не соответствующим конкретной ситуации, а иногда и вовсе ошибочным. То же касается жестов и мимики. Отдельной проблемой современных аудиовизуальных систем синтеза является несоответствие между звуковыми и паралингвистическими характеристиками, в том числе: неточная или неправильная синхронизация, несоответствие семантике произно-

симого текста и общей эмоциональной окраске, совмещение несочетаемых жестов, мимики и интонации. При этом даже небольшого отклонения от естественной речи, как правило, достаточно для того, чтобы возник широко известный эффект «зловещей долины». В связи с этим особое значение приобретает рассмотрение вопросов автоматического анализа мимики в сопоставлении с анализом интонационных характеристик в эмоционально-окрашенных высказываниях.

На начальном этапе исследования были рассмотрены существующие математические модели анализа мимики по изображению. Общим в данных моделях является наличие предварительной обработки изображений, включающее нахождение области лица, обрезку и масштабирование найденной области, выравнивание лица, регулировку контрастности и т.п.; затем производится извлечение визуальных признаков с помощью различных методов (на основе моделей внешнего вида, на основе глобальных и локальных объектов либо на основе геометрических объектов); и, наконец, происходит формирование обобщающих выводов (например, в виде классификации эмоций на основе предложенной П. Экманом и У. Фризеном [1] эмоциональной системы кодирования лицевых движений).

В качестве инструмента анализа нами был выбран Python Facial Expression Analysis Toolbox (Py-Feat) [2], состоящий из различных моделей, которые позволяют извлекать черты лица, а также моделей предварительной обработки и визуализации полученных данных в виде графиков. Использование пакета Py-feat для проведения классификации эмоций на базе предварительно обученной нейронной сети показал хорошие результаты при анализе видеозаписей американского варианта английской речи, в том числе на материале кинопродукции. Однако в ходе пилотного эксперимента по применению данного пакета для анализа мимики носителей русского языка нами были обнаружены ошибки в трактовке эмоций. В связи с этим было решено самостоятельно расширить функционал Py-feat, разработав и обучив собственную нейронную сеть на материале русского языка.

Эффективное решение подобной задачи мог обеспечить лишь такой мультимедийный корпус, который включал бы в себя: (а) оценку эмоционального значения носителями языка; (б) не только орфографическую расшифровку, но и подробную фонетическую аннотацию звукового сигнала, содержащую в том числе сведения о границах речевых единиц. Разработанный ранее на кафедре фонетики и методики преподавания иностранных языков Санкт-Петербургского государственного университета корпус русской иронической речи полностью отвечал данным критериям. Подробная фонетическая аннотация звукового сигнала в данном корпусе позволила провести анализ интонационных характеристик в сопоставлении с анализом мимики в ироничных и неироничных высказываниях. Таким образом, несмотря на то, что ирония не является эмоцией как таковой, но понимается

большинством авторов как эмоционально-оценочная коннотация, данный корпус позволил протестировать инструмент Py-Feat для автоматического определения эмоционально-оценочного значения.

Мультимедийный корпус русской иронической речи (подробное описание см. в [3, 4]) построен таким образом, чтобы обеспечить сравнение нейтральных и ироничных фрагментов с идентичным лексическим составом. Для этого были составлены контексты, предполагающие прочтение с иронией или без иронии. При этом ремарки (например, *сказал он иронично*, *спросила она язвительно*), эксплицитно указывающие на ироническое значение, отсутствовали. В эти контексты были помещены омонимичные целевые фрагменты. Например, один и тот же целевой фрагмент *русалочка* был включен в разные контексты: (а) *Чуть не утонула! Русалочка! Ты зачем туда поплыла?* и (б) – *Какой твой любимый персонаж детских сказок? – Русалочка!*. Целевые фрагменты состояли из одной синтагмы.

Дикторы читали целиком представленные контексты (короткие монологи и диалоги, состоящие из 2–4 фраз) вместе с целевым фрагментом. Дикторов просили прочитать естественно, так, как они бы произнесли данные высказывания в обычной разговорной речи. Для проверки данных, полученных на материале анализа коротких текстов, были составлены и более развернутые связные тексты, включающие в себя ироничные целевые фрагменты.

Аудиозапись производилась с использованием профессионального звукозаписывающего оборудования и программного обеспечения Nuendo с частотой дискретизации 44100 Гц в акустической кабине кафедры фонетики и методики преподавания иностранных языков Санкт-Петербургского государственного университета. Параллельно с аудиозаписью проводилась видеозапись на видеокамеру Sony Handycam FDR-AX700 со скоростью 100 кадров в секунду.

Из полученных аудиозаписей были вырезаны ироничные и неироничные целевые фрагменты, которые предъявлялись вне контекста слушателям в ходе аудиторского эксперимента. Участники эксперимента должны были сопоставить звучащий целевой фрагмент с одним из представленных на экране мини-текстов, опираясь лишь на звучание отрывка. Таким образом были получены оценки наличия или отсутствия ироничного значения в целевых фрагментах носителями языка. Далее проводилась подробная фонетическая аннотация, включающая определение границ фонетических единиц разного уровня.

Аннотированные аудиозаписи были синхронизированы с видеозаписями, что позволило сопоставить реализацию мимических движений с границами целевых фрагментов и контекстов в целом. Поскольку материал для чтения был представлен дикторам в распечатанном виде, набор жестовых и мимических движений был ограничен: лишь некоторые дикторы использовали движения руками, в редких случаях дикторы отрывали взгляд от распечатанного материала. В связи с этим в настоящем исследовании

рассматриваются единицы движения (action units), описывающие движения бровей: AU1 – подниматель внутренней части брови, AU2 – подниматель внешней части брови, AU4 – опускатель брови. Еще одной причиной выбора этих параметров является тот факт, что они не зависят от артикуляции при произнесении в отличие, например, от единиц действия, отвечающие за движения губ. Кроме того, именно движение бровей отмечается исследователями как наиболее яркая мимическая характеристика иронической речи в работе на материале актерской речи у носителей американского варианта английского языка [5]. Можно предположить, что и в русском языке движение бровей будет коррелировать с присутствием иронического значения в высказывании.

Для анализа аудио- и видеосигнала были выбраны целевые фрагменты, прочитанные без иронии и с иронией, а также были рассмотрены контексты (мини-тексты), включающие данные целевые фрагменты. При сравнении как целевых фрагментов, так и окружающих их контекстов, были обнаружены парадигматические отличия речи без иронии и речи с иронией как на уровне акустических характеристик, так и на уровне мимики (в частности, движения бровей). На иллюстрации ниже (Рис. 1) можно заметить б*о*льшую амплитуду мимических движений для всех проанализированных единиц движения (AU 1, AU 2, AU 4) в восклицательном высказывании без иронии *Ты так блестяще выступил. Какой молодец!* и меньшую – в высказывании с иронией *Ушел и никого не предупредил. Какой молодец!*. Причем это характерно, как уже говорилось выше, не только для целевых фрагментов (они выделены на графиках заливкой), но и для всего высказывания (контекста) в целом.

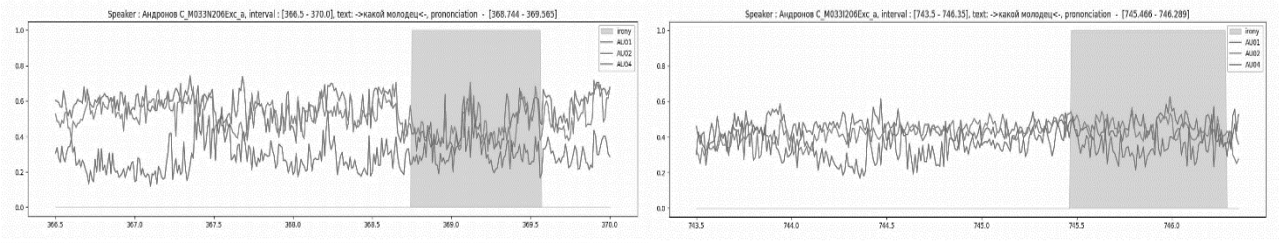


**Рис. 1.** Графики изменения положения в пространстве единиц движения (action units), отвечающих за движения бровей в высказывании без иронии *Ты так блестяще выступил. Какой молодец!* (слева) и *Ушел и никого не предупредил. Какой молодец!*; целевой фрагмент *Какой молодец!* выделен заливкой; мужская речь, диктор М033.

Не менее яркие отличия присутствуют и в аудиосигнале. На рисунке 2 представлены графики ОТ для восклицательного высказывания без иронии и высказывания с иронией. При выражении иронии наблюдается увеличение мелодического диапазона как в целевом фрагменте, так и во всем контексте. Кроме того, увеличивается длительность ударных слогов и мелодический интервал внутри ударного слога.

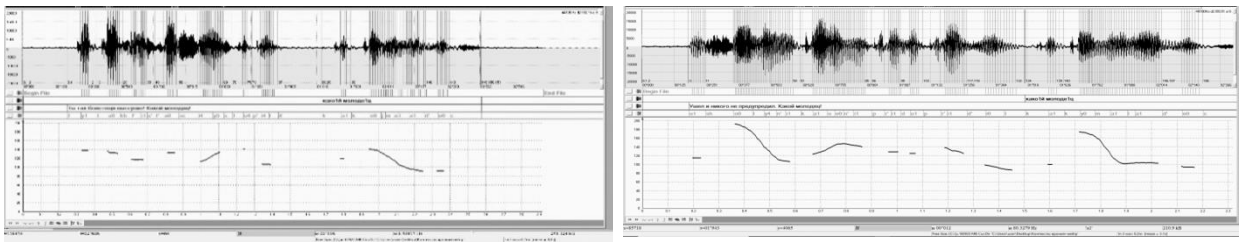

**Рис. 2.** Графики ОТ в высказывании без иронии *Ты так блестяще выступил. Какой молодец!* (слева) и *Ушел и никого не предупредил. Какой молодец!*; мужская речь, диктор М033.

Таким образом полученные результаты свидетельствуют не только о наличии контраста между иронической и неиронической речью одновременно на акустическом и на паралингвистическом уровнях, но также позволяют предположить, что выражение определенной эмоции или эмоционально-оценочной коннотации, по-видимому, происходит не только внутри целевой синтагмы, но и за ее рамками. Подобные выводы о порождении речевого высказывания на различных уровнях могут быть учтены разработчиками при создании и усовершенствовании диалоговых систем с использованием технологий искусственного интеллекта.

## ЛИТЕРАТУРА


1. *Ekman P., Friesen W.* Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto : Consulting Psychologists Press, 1978.
2. https://py-feat.org/ (дата обращения: 04.04.2024)
3. Can We Detect Irony in Speech Using Phonetic Characteristics Only? Looking for a Methodology of Analysis / P. *Skrelin, U. Kochetkova, V. Evdokimova, D. Novoselova* // SPECOM 2020. Singapore : Springer Nature, 2020. P. 544–553.
4. The Multimedia Corpus of Russian Ironic Speech for Phonetic Analysis / U. *Kochetkova, P. Skrelin, V. Evdokimova, T. Kachkovskaia* // Literature, Language and Computing. Singapore : Springer Nature, 2023. P. 223–237.
5. *Tabacaru S., Lemmens M.* Raised eyebrows as gestural triggers in humour: The case of sarcasm and hyper-understanding // The European Journal of Humour Research. 2014. Vol. 2. P. 11–31.