УДК 57.087.1:004.934.2 DOI: 10.18698/2541-8009-2018-6-334

ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ СТИМУЛЬНОГО МАТЕРИАЛА ДЛЯ ФОРМИРОВАНИЯ БАЗЫ ДАННЫХ ЭМОЦИОНАЛЬНО ОКРАШЕННОЙ РЕЧИ

А.В. Хорава albinahorava@mail.ru
SPIN-код: 8009-3680

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Звучащая речь является одним из видов сигналов, используемых человеческим мозгом для анализа эмоционального состояния человека. В настоящее время активно развивается направление распознавания эмоций по речи с помощью компьютерных систем. Результат работы алгоритмов распознавания эмоций по речи во многом определяется базой, которая применяется для обучения алгоритмов. Общедоступная база данных эмоционально окрашенной русской речи в настоящее время отсутствует. В данной работе предпринята попытка устранения указанного недостатка. Описан стимульный материал для индуцирования эмоций говорящего. Приведены параметры отдельных стимулов (текста и видеозаписи), использующихся в процессе формирования базы.

Ключевые слова

Базовые эмоции, характеристики сигнала, интенсивность речевого сигнала, частота основного тона, частоты первых формант, темп речи

Поступила в редакцию 26.04.2018

Введение. Человеческий мозг распознает эмоции другого человека по мимике, жестам и речи [1, 2]. В данной статье рассмотрены нейтральное состояние (англ. *neutral*) и шесть базовых эмоций в соответствии с классификацией, предложенной П. Экманом и У. Фризеном:

– радость (англ. *happiness*);
– печаль (англ. *sadness*);
– гнев (англ. *anger*);
– удивление (англ. *surprise*);
– страх (англ. *fear*);
– отвращение (англ. *disgust*).

Характеристики речевого сигнала, использующиеся для определения эмоций по речи. Для решения задач определения эмоционального состояния человека по речи используют следующие численные характеристики, основанные на количественной оценке паралингвистических характеристик речевого сигнала [3, 4]:

1) энергию (интенсивность) речевого сигнала;

2) частоты первых формант речевого сигнала и их среднеквадратические отклонения (СКО);

3) частоту основного тона (ЧОТ) и его СКО;

4) изменение темпа речи (растяжение-сжатие речевого сигнала во времени, табл. 1).

Таблица 1

Паралингвистические характеристики речевого сигнала

Характеристика речевого сигнала	Описание характеристики
Громкость (интенсивность) речи	Шепот (менее 20 дБ) Значительное снижение (20…40 дБ) Умеренное понижение (40…50 дБ) 50…80 дБ (при постоянном фоновом шуме до 10 дБ) Умеренное повышение (80…90 дБ) Значительное повышение (90…110 дБ) Крик (свыше 110 дБ)
Формантные частоты	Частота максимальной амплитуды в пределах форманты
Частота основного тона	Минимальная длина фрагмента звучащей речи 100…500 мс С вероятностью 0,95 основной тон мужских голосов расположен в интервале 97…195 Гц, женских голосов — в интервале 195…320 Гц
Темп речи	Зависит от индивидуальных особенностей диктора
Паузация	Короткие паузы — до 3 с Средние паузы — 3…7 с Длинные паузы — свыше 7 с

По результатам исследований, проведенных Центром речевых технологий [5], определены наиболее часто встречающиеся относительные изменения характеристик речевого сигнала для каждой базовой эмоции (табл. 2).

Таблица 2

Изменение паралингвистических характеристик речевого сигнала

Изменение значений относительно нейтрального состояния	Базовая эмоция				
	Радость	Печаль	Гнев	Удивление	Страх
Изменение количества пауз	↓	↑	↓	↑	↑
Изменение длительности пауз	↓	↓	↓	↑	↑
Изменение ЧОТ	↑	↓	↓	↓	↑
Изменение СКО ЧОТ	↑	↓	↑	↑	↑
Изменение частоты 1-й форманты	↑	↓	↓	↓	↓
Изменение частоты 2-й форманты	↑	↓	↓	↓	↓
↑ — значение повысилось относительно нейтрального состояния; ↓ — значение понизилось относительно нейтрального состояния.					

Существующие базы данных речевых сигналов. Для определения эмоционального состояния человека по речи могут быть использованы различные виды классификаторов. При этом результат работы алгоритмов распознавания эмоций в значительной степени определяется базой данных речевых сигналов, на которой осуществлялось обучение классификаторов.

Существующие базы данных речевых сигналов, как правило, содержат речь на английском или немецком языках, например:

1) FAU AIBO — база данных спонтанных эмоций. Набор данных состоит из девяти часов речи на немецком языке от 51 ребенка в возрасте от 10 до 13 лет при их взаимодействии с домашним животным-роботом. Каждый аудиофайл состоит из одного короткого предложения. Каждая аудиозапись отнесена к одной из пяти категорий: гнев, выразительность, нейтральное состояние, положительная эмоция, скука;

2) SAVEE — база данных на английском языке. Набор данных состоит из речи четырех актеров мужского пола. Эмоциональные типы для каждого высказывания соответствуют одной из шести базовых эмоций (радость, печаль, гнев, удивление, страх, отвращение) или нейтральному состоянию;

3) VAM — база данных, созданная в университете Карлсруэ. Состоит из высказываний, полученных из популярного немецкого ток-шоу. Эмоциональные типы базы: счастье-интерес, сердится-беспокоится, грустно-скучно, расслабленный-спокойный. Аудиозаписи базы данных содержат реальные (не индуцированные) эмоции [6];

4) Berlin — база данных, созданная в Техническом университете Берлина и состоящая из эмоциональных высказываний на немецком языке, произнесенных десятью актерами, среди которых пять мужчин и пять женщин. Каждое высказывание имеет один из следующих эмоциональных типов: пять базовых эмоций (радость, печаль, гнев, страх, отвращение), скука и нейтральное состояние [7].

Описание процесса сбора базы данных эмоционально окрашенной русской речи. При формировании эмоционально окрашенной базы необходимо учитывать следующее [8]:

– большое количество эмоциональных типов снижает качество распознавания;

– эмоциональные типы должны быть ярко отличимыми друг от друга, а также успешно и достоверно определяться на слух подавляющим большинством экспертов.

Для создания базы данных эмоционально окрашенной русской речи были выбраны два типа стимульного материала: текст и видеозаписи. С их помощью лучше индуцировалась та или иная эмоция. Если эмоцию не удавалось вызвать путем прочтения фрагмента текста, демонстрировали видеозапись. Аудиозаписи не использовали в качестве стимульного материала, поскольку зачастую у диктора не выявлялся эмоциональный отклик.

Сбор осуществляли по методике формирования баз данных SAVEE и Berlin. Были рассмотрены шесть базовых эмоций и нейтральное состояние. Таким образом, общее количество аудиозаписей для каждого диктора составило семь.

Аудиорегистрацию проводили в тихом помещении с помощью компьютерного микрофона и ноутбука. Параметры полученных аудиозаписей приведены ниже:

– формат аудиофайла — wav;

– тип кодирования — импульсно-кодовая модуляция (ИКМ; англ. *Pulse Code Modulation*, PCM);

– количество каналов — 2 (аудиозапись осуществлялась в режиме «моно», значения для двух каналов одинаковы);

– глубина кодирования — 16 бит/канал;

– частота дискретизации — 48,0 кГц;

– скорость потока данных — 1 536 Кбит/с.

При записи сигналов в качестве приоритетного выступало индуцирование эмоций с помощью воспоминаний из реальной жизни диктора. Таким образом регистрировались базовые эмоции «гнев», «страх» и «отвращение». Для регистрации других базовых эмоций применяли стимульные материалы двух основных типов:

– фрагменты текстов художественных произведений;

– видеозаписи.

При этом фрагменты текстов художественных произведений и видеозаписи подбирали таким образом, чтобы их содержание вызывало реакцию, соответствующую заданной эмоции. В зависимости от типа индуцируемой эмоции осуществляли параллельную или последовательную регистрацию.

При последовательной регистрации диктор пересказывал содержимое просмотренной видеозаписи через 20 с после ее завершения. Таким образом осуществлялась регистрация аудиозаписей, соответствующих базовой эмоции «печаль».

При параллельной регистрации диктор прочитывал фрагмент текста художественного произведения (для базовых эмоций «радость» и «печаль» и нейтрального состояния) или пересказывал содержимое видеозаписи одновременно с ее просмотром (для базовой эмоции «удивление»). Суть данного типа регистрации состоит в дополнительном воздействии на эмоциональное состояние человека и одновременной регистрации аудиосигнала [9].

Отметим, что для регистрации базовой эмоции «печаль» были использованы оба типа стимульного материала (тексты и видеозаписи). Если эмоцию не удавалось вызвать путем прочтения фрагмента текста, осуществляли демонстрацию видеозаписи с последующей аудиорегистрацией.

Базовые эмоции, тип используемого стимульного материала и тип регистрации представлены в табл. 3.

Таблица 3

Стимульный материал, использованный для формирования базы данных

Базовая эмоция	Тип стимульного материала	Тип регистрации
Радость	Фрагмент текста	Параллельная
Печаль	Фрагмент текста	Параллельная
	Видеозапись	Последовательная
Гнев	Не использовался	—
Удивление	Видеозапись	Параллельная
Страх	Не использовался	—
Отвращение	Не использовался	—
Нейтральное состояние	Фрагмент текста	Параллельная

Еще раз подчеркнем, что отрицательные эмоции («гнев», «страх» и «отвращение») индуцируются проще, чем положительные («радость», «удивление») или нейтральное состояние, поэтому при их регистрации стимульный материал не использовался.

В сборе базы данных принимали участие 20 человек (10 женщин и 10 мужчин) в возрасте от 18 до 25 лет. Для каждого диктора были зарегистрированы по семь аудиозаписей (соответствующих шести базовым эмоциям и нейтральному состоянию), длительность каждой аудиозаписи составила около 1 мин. В табл. 4 приведены длительности аудиозаписей для дикторов мужского пола, а в табл. 5 — для дикторов женского пола. Общее количество зарегистрированных аудиозаписей — 140. Общая продолжительность аудиозаписей базы составила порядка 2,5 ч.

Таблица 4

Длительность аудиозаписей для дикторов мужского пола (минуты: секунды)

Диктор	Базовая эмоция						
	Радость	Печаль	Гнев	Удивление	Страх	Отвращение	Нейтральное состояние
m_01	00:57	00:59	01:26	00:43	00:28	00:57	01:03
m_02	00:43	00:49	00:40	00:44	00:27	00:37	01:03
m_03	00:38	00:36	00:39	00:41	00:52	00:37	00:46
m_04	00:44	00:41	00:41	00:53	00:57	00:41	00:39
m_05	00:47	00:53	00:44	00:43	00:47	00:45	00:44
m_06	00:51	00:55	00:45	00:33	00:40	00:53	00:58
m_07	00:43	00:40	00:39	00:41	00:59	00:46	00:41
m_08	00:37	00:35	00:40	00:44	00:39	00:42	00:50
m_09	00:46	00:54	00:54	00:44	00:44	00:51	00:45
m_10	00:46	00:53	00:50	00:53	00:47	00:45	00:50

Таблица 5

Длительность аудиозаписей для дикторов женского пола (минуты: секунды)

Диктор	Базовая эмоция						
	Радость	Печаль	Гнев	Удивление	Страх	Отвращение	Нейтральное состояние
f_01	01:02	01:05	00:45	01:02	00:41	00:51	01:00
f_02	02:29	01:08	01:25	01:49	00:30	01:23	01:04
f_03	01:11	01:11	01:00	00:54	00:48	01:10	00:58
f_04	00:58	01:12	01:03	00:39	00:39	00:59	01:09
f_05	00:56	01:08	01:03	00:56	00:51	01:05	01:01
f_06	00:50	00:58	00:45	00:43	00:45	01:06	01:00
f_07	01:10	01:10	00:59	00:58	00:55	00:55	00:51
f_08	01:15	01:02	01:01	00:52	00:58	00:45	01:00
f_09	01:00	00:45	00:45	00:38	01:00	01:01	01:20
f_10	00:55	01:00	01:04	00:41	00:47	00:56	01:05

Таким образом, на основе созданной базы, в которой индуцировались шесть эмоций (радость, печаль, гнев, удивление, страх, отвращение) и нейтральное состояние, можно разработать алгоритм классификации психоэмоционального состояния (ПЭС) человека. Метод обработки речевого сигнала при анализе ПЭС является достаточно удобным, поскольку для его функционирования необходим только микрофон.

Литература

[1] Бойко А.А., Неверова Е.С., Каранкевич А.И., Спиридонов И.Н. Исследование невербального поведения студентов при сдаче экзаменов. *Наука и инженерное образование. SEE-2016*. Москва, 2016, с. 162–163.

[2] Пилипенко М.Н., Латышева Е.Ю., Бойко А.А., Спиридонов И.Н. Исследование алгоритмов автоматического обнаружения двигательных единиц по изображению лица. *Биотехносфера*, 2016, № 6(48), с. 8–12.

[3] Кипяткова И.С., Карпов А.А. Аналитический обзор систем распознавания русской речи с большим словарем. *Труды СПИИРАН*, 2010, № 1(12), с. 7–20.

[4] Стерлинг Г.Г., Приходько П.В. Глубокое обучение в задаче распознавания эмоций из речи. *Сб. тр. 40 междисциплинарной школы-конф. «Информационные технологии и системы 2016»*. Москва, ИППИ РАН, 2016, с. 451–456.

[5] Центр речевых технологий. URL: http://www.speechpro.ru/ (дата обращения 19.11.2017).

[6] Алешин Т.С., Редько А.Ю. Принципы подготовки баз речевых данных для задачи распознавания эмоционального окраса речи человека по речевому сигналу. *Современные наукоемкие технологии*, 2016, № 6-2, с. 229–234.

[7] Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B. A database of German emotional speech. *Proc. Interspeech*, 2005, pp. 1517–1520.

[8] Давыдов А.Г., Киселев В.В., Кочетков Д.С. Классификация эмоционального состояния диктора по голосу: проблемы и решения. *Тр. межд. конф. «Диалог – 2011»*. Москва, РГТУ, 2011, с. 178–185.

[9] Изард К. *Психология эмоций*. Санкт-Петербург, Питер, 2000, 464 с.

Хорава Альбина Вахуштьевна — магистрант кафедры «Биомедицинские технические системы», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Бойко Андрей Алексеевич, ассистент кафедры «Биомедицинские технические системы», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

PECULIAR FEATURES OF USING THE STIMULUS MATERIAL FOR CREATING THE EMOTIONALLY COLOURED SPEECH DATABASE

A.V. Horava
albinahorava@mail.ru
SPIN-code: 8009-3680

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

The sounding speech is one of the kinds of signals used by the human brain for analyzing the emotional state of a person. Emotion recognition from speech with the aid of computer systems is currently an actively developing line of research. Emotion recognition from speech algorithm output is mainly determined by the base used for teaching algorithms. At present there is no public database of the emotionally coloured Russian speech. In this paper we try to remove the specified shortcoming. The article describes the stimulus material for inducing the speaker's emotions. We provide the parameters of separate stimuli (text and video recording), used in the process of forming the base.

Keywords

Basic emotions, signal characteristics, speech signal intensity, base frequency, first formant frequencies, tempo of speech

Received 26.04.2018

References

[1] Boyko A.A., Neverova E.S., Karankevich A.I., Spiridonov I.N. Issledovanie neverbal'nogo povedeniya studentov pri sdache ekzamenov [Research on students nonverbal behavior at the exam]. *Nauka i inzhenernoe obrazovanie. SEE-2016* [Science and engineering education. SEE–2016]. Moscow, 2016, pp. 162–163.

[2] Pilipenko M.N., Latysheva E.Yu., Boyko A.A., Spiridonov I.N. Research of algorithms for action units' automatic detection using facial image. *Biotekhnosfera*, 2016, no. 6(48), pp. 8–12.

[3] Kipyatkova I.S., Karpov A.A. An analytical survey of large vocabulary Russian speech recognition systems. *Trudy SPIIRAN* [SPIIRAS Proceedings], 2010, no. 1(12), pp. 7–20.

[4] Sterling G.G., Prikhod'ko P.V. Glubokoe obuchenie v zadache raspoznavaniya emotsiy iz rechi [Deep learning in problem of emotion recognition from speech]. *Sb. tr. 40 mezhdistsiplinarnoy shkoly-konf. "Informatsionnye tekhnologii i sistemy 2016"* [Proc. 40th Interdisciplinary Conf. & School "Information Technology and Systems 2016"]. Moscow, IITP RAS, 2016, pp. 451–456.

[5] Tsentr rechevykh tekhnologiy [Center of speech technologies]. Available at: http://www.speechpro.ru/ (accessed 19 November 2017).

[6] Aleshin T.S., Red'ko A.Yu. Bases of speech data corpus preparation for the emotional speech recognition. *Sovremennye naukoemkie tekhnologii* [Modern high technologies], 2016, no. 6-2, pp. 229–234.

[7] Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B. A database of German emotional speech. *Proc. Interspeech*, 2005, pp. 1517–1520.

[8] Davydov A.G., Kiselev V.V., Kochetkov D.S. Klassifikatsiya emotsional'nogo sostoyaniya diktora po golosu: problemy i resheniya [Classification of speaker emotional state by voice: problems and solutions]. *Tr. mezhd. konf. "Dialog – 2011"* [Proc. Int. conf. "Dialogue-2011"]. Moscow, RGTU, 2011, pp. 178–185.

[9] Izard C.E. The psychology of emotions. Springer Science & Business Media, 1991, 452 p. (Russ. ed.: Psikhologiya emotsiy. Sankt-Petersburg, Piter publ., 2000, 464 p.)

Horava A.V. — Master's Degree student, Department of Biomedical Engineering Systems, Bauman Moscow State Technical University, Moscow, Russian Federation.

Scientific advisor — A.A. Boyko, Assistant, Department of Biomedical Engineering Systems, Bauman Moscow State Technical University, Moscow, Russian Federation.