УДК 004.934

ПРИНЦИПЫ ПОДГОТОВКИ БАЗ РЕЧЕВЫХ ДАННЫХ ДЛЯ ЗАДАЧИ РАСПОЗНАВАНИЯ ЭМОЦИОНАЛЬНОГО ОКРАСА РЕЧИ ЧЕЛОВЕКА ПО РЕЧЕВОМУ СИГНАЛУ

Алешин Т.С., Редько А.Ю.

Санкт-Петербургский государственный институт кино и телевидения, Санкт-Петербург, e-mail: mar-len90@mail.ru, redkant@yandex.ru

Аффективные вычисления – это одна из самых развивающихся областей в сфере исследований взаимодействия человека и вычислительных машин. Одной из составных его частей является распознавание эмоций по речи человека. На сегодняшний день представлено множество алгоритмов распознавания эмоций, однако пока не существует общих требований для составления обучающей и тренировочной базы данных. В данной работе рассмотрены виды баз данных, эмоциональные типы и их количество, применяемые для задач распознавания эмоций по речевому сигналу человека, а также предложен вариант составления универсальной базы данных для обучения и тренировки алгоритма распознавания. Интересной задачей при распознавании является выбор и построение классификатора для получения наиболее высокого процента точности распознавания эмоций.

Ключевые слова: речевой сигнал, эмоционально окрашенная речь, распознавание эмоций по речевому сигналу, эмоциональные признаки речи

BASES OF SPEECH DATA CORPUS PREPARATION FOR THE EMOTIONAL SPEECH RECOGNITION

Aleshin T.S., Redko A.Yu.

Saint-Petersburg State University of Film and Television, Saint-Petersburg, e-mail: mar-len90@mail.ru, redkant@yandex.ru

Affective computing is one of the fastest growing areas in the field of studies of the man and computers interaction. Human voice emotion recognition is a component of it. Today are a lot of emotion recognition algorithms but there are no general requirements for the preparation of teaching and training databases so far. This paper discusses types of databases, emotional types and the number used for emotion recognition tasks of human speech signal. Authorsnoffered the option of drawing up a universal database for learning and training recognition algorithm. The selection and construction of the classifier is an interesting challenge in recognizing to produce the highest percentage of accuracy of emotion recognition.

Keywords: speech signal, emotional speech, recognition, emotional speech features

Аффективные вычисления – это технологии, обеспечивающие взаимодействие между человеком и компьютером с помощью эмоций. Например, компьютер распознаёт эмоциональное состояние пользователя и на этой основе принимает то или иное решение. В такой цепочке ключевым звеном является процесс распознавания эмоций человека. Человеческий мозг распознаёт эмоции другого человека по мимике, жестам и речи. Однако составить алгоритм, позволяющий распознать эмоции человека по этим трём видам коммуникации, чрезвычайно сложно.

В данной работе рассмотрим распознавание эмоционально окрашенной речи человека по его речевому сигналу. Обзор работ по данной теме показал, что выявлено множество признаков в речевом сигнале, по которым можно определить тот или иной эмоциональный окрас речи [1], также для задачи распознавания эмоций применяются различные виды классификаторов. Однако результат работы алгоритмов распознавания эмоций во многом зависит от базы данных, содержащей наборы речевых сигналов.

Также в различных работах выделяют различное количество типов эмоций: от двух [5] до пятнадцати [4].

Рассмотрим некоторые исследования и сделаем обобщение относительно того, какой должна быть база данных для обучения и тестирования алгоритма распознавания эмоций по эмоционально окрашенной речи.

Распознавание эмоций в речи является сложной задачей, потому что нет однозначного ответа на то, какая эмоция «правильная» для данного слова или выражения. Также возникают трудности с определением спонтанных эмоций, так как по факту эмоции изменяются динамически и определить каждую с уверенностью в определенный момент времени достаточно сложно, а порой и невозможно. Кроме того, эмоции, которые выражаются в спонтанной речи, труднее поддаются распознаванию, нежели эмоции в речи человека, выступающего на сцене.

Распознавание эмоций речи из записей телефонных разговоров

В работе [6] представлен результат исследования, в котором были использованы данные коммерческого центра обработки вызовов. Как это часто бывает, при определении эмоций по акустическим характеристикам лишь немногие отличались алгоритмом распознавания от нейтральной.

В спонтанной речи появление канонических эмоций, таких как счастье и гнев, как правило, маловероятно. Если использовать в качестве обучающих данных базу, составленную из записей телефонных разговоров, то распределение данных по эмоциональным классам будет весьма несбалансированным, что усложняет задачу распознавания.

Материал, используемый в данной работе, был записан в шведской Компании Голосовых Услуг (КГУ). Высказывания разделены опытным старшим научным сотрудником компании на: нейтральную, акцентированную и отрицательную (разочарование) речь. Подмножество материала классифицировали с помощью пяти различных лиц и пар. Как показало тестирование, большинство высказываний являются нейтральными (не выразительными, ≈ 93 %), но несколько процентов абонентов были определены как разочарованные (≈ 5 %), оставшиеся 2 % пришлись на акцентированную речь.

Такой процент распознавания связан именно с неравномерностью обучающей базы данных, а также с особенностью проведения эксперимента.

Распознавание эмоций по слоговым компонентам данных баз, предназначенных для распознавания речи

Далее рассмотрим работу [3], в которой анализируются две базы данных на немецком языке: база данных Ollo, предназначенная для проведения экспериментов по распознаванию речи на речевой изменчивости, и эмоциональная база Берлин, предназначенная для анализа и синтеза эмоциональной речи. В статье делается попытка определить отношение между внутренней речевой изменчивостью и эмоциями и рассматривается это соотношение с точки зрения распознавания речи.

Акустический анализ проводится на обеих базах данных, проанализированы гласные [a], [e], [i], [o] и [u], включая их длинные и короткие варианты.

База данных Берлин состоит из десяти высказываний, которые производились пятью мужчинами и пятью женщинами, немецкими актерами. Высказывания были записаны по типовому закону эмоций, а также был записан и нейтральный вариант. Предложения были взяты из повседневного общения и могут быть интерпретированы в контексте эмоций, но они не содержат смысловой (семантической) нагрузки.

В итоге эксперимент по распознаванию эмоционального окраса речи отдельно для мужского голоса и для женского дал результат, представленный в табл. 1 и 2. В этих таблицах показана связь между эмоциональными аспектами и речевой изменчивостью, которые оцениваются отдельно для женщин и мужчин.

Таблица 1

Связь эмоциональных аспектов и вариабельности речи для женщин

Изменения речи	Эмоциональный аспект
Быстро, громко, тихо	Счастье
Медленно	Печаль
Вопросительная интонация	Беспокойство, страх
Стиль изложения	Нейтраль

Таблица 2

Связь эмоциональных аспектов и вариабельности речи для мужчин

Изменения речи	Эмоциональный аспект
Быстро, громко, тихо	Скука
Медленно	Нейтраль
Вопросительная интонация	Скука
Стиль изложения	Нейтраль

Существует распространенное убеждение, что мужчины и женщины выражают свои эмоции по-разному. Данное исследование показывает, что эмоции счастья, печали и страха более характерны для женщин. Исследования также показывают, что мужчины выражают меньше тревоги и депрессии, чем женщины.

В этой работе не был назначен такой эмоциональный аспект, как «нормальный», для изменчивости речи женщин, так как ни одно из расчетных расстояний до векторов эмоциональных признаков не было достаточно низким.

Результат эксперимента оказался любопытным. Однако мужская речь действительно более скупа на эмоции, нежели женская. К тому же мгновенная частота мужского и женского речевого сигнала будет различна, поэтому уместно классифицировать эмоции в женском голосе и мужском раздельно.

База данных эмоциональной речи, составленная из речи детей

Следующей работой, рассмотренной нами, является статья [2]. Здесь эксперимент по распознаванию эмоций проводится на FAU AIBO – базе данных спонтанных эмоций.

FAU AIBO – набор данных, содержащих эмоциональную речь. Набор данных состоит из записей разговоров немецких детей, взаимодействующих с домашним животным-роботом. База состоит из 9959 отрезков для обучения и 8257 отрезков для тестирования. Отрезки оценивались по пяти категориям эмоций: злость (*A*), выразительность (E), нейтраль (*N*), положительность (P, состоящий из детских интонаций и радости) и отдых (R, состоящий из эмоций, не принадлежащих к другим категориям, таким как скука, беспомощность и так далее). Распределение по пяти классам весьма несбалансированное. Например, процент обучающих данных из каждого класса выглядит следующим образом: *A* (8,8 %), E (21 %), *N* (56,1 %), *P* (6,8 %), *R* (7,2 %).

Максимальный процент распознавания алгоритма, представленного в [2], это 44 % распознавания. Процент невысок, это связано с особенностями выбора признаков и классификаторов. Интерес вызывает технологичность и необычный подход к сбору материала для базы данных. Кроме того, эмоции детей всегда яркие и выразительные. Однако сложно получить все виды эмоциональных реакций от детей. К примеру, эмоцию печаль получить достаточно сложно. Заметим также неточную задачу эмоциональных типов. Скажем, тип «выразительность» был получен гораздо чаще, чем положительность. Хотя реакция детей на игрушку-робота должна быть смещена в сторону положительности. Да и трудно оценить такой эмоциональный тип, как положительность. Это может быть например как удивление, так и счастье, которые в данной классификации могут относиться и к выразительности.

Распознавание 2, 4, 6 и 15 эмоций по базе данных стихотворных отрывков и стенограмм

Наиболее интересными и разнообразными с точки зрения итогового результата, оказались исследования [4]. В данной исследовательской работе были проанализированы 2442 высказывания из эмоционального прочтения стихов и стенограмм и выделено 62 признака из каждого высказывания.

В итоге получено более 90 % точности в распознавании гнева и нейтральной эмоции, более 80 % точности в выделении счастья и печали, а также в отличии горячей и холодной злости. Также достигнуты 62 и 49 % точности классификации для 4 и 6 эмоций соответственно. И кроме того, был получено 20 % точности в классификации всех 15 эмоций в базе, что является большим прорывом по сравнению с другими исследованиями.

Представленный в [4] алгоритм распознавания эмоций по динамике речи не зависит.

В группе из четырёх эмоций выделены следующие типы: счастье, печаль, гнев, нейтраль. Из шести эмоций выделены счастье, печаль, гнев, нейтраль, интерес и паника.

Среди 15 типов эмоций указаны: нейтральные, отвращение, паника, тревога, горячий гнев, холодный гнев, отчаяние, печаль, восторг, счастье, интерес, скука, стыд, гордость и презрение.

Стоит заметить, что для обучения и тестирования данного алгоритма использовалась база данных из записей профессиональных актёров, как мужчин, так и женщин.

Распознавание эмоций по базам эмоциональной речи на английском и немецком языках

А теперь перейдём к исследованию [7], в котором использовались эмоциональные речевые базы данных на 2-х языках (английский, немецкий).

База эмоциональных данных Берлин была записана в Техническом университете Берлина и состоит из эмоциональных высказываний на немецком языке, произнесённые 10 актерами (среди них 5 женщин).

Каждое высказывание имеет один из следующих эмоциональных типов: нейтральный, гнев, страх, радость, печаль, скука или отвращение.

SAVEE (Аудио-, видеовыражения эмоций из Суррей) – база данных, записанная как часть исследования по аудиовизуальной классификации эмоций четырьмя мужчинами, носителями английского языка. Эмоциональные типы для каждого высказывания базы данных SAVEE: гнев, отвращение, страх, счастье, печаль, удивление и нейтральное.

Кроме того, в исследовании использовалась база данных VAM, созданная в университете Карлсруэ. Она состоит из высказываний, полученных из популярного немецкого ток-шоу «Вера во второй половине дня». Эмоциональные типы базы: счастье-интерес, сердится-беспокоится, грустно-скучно, и расслабленный-спокойный.

Две базы (Берлин, SAVEE) состоят из профессионально отработанных эмоций, в то время как база данных VAM включает в себя реальные.

По сравнению с базами данных Берлин и SAVEE, база данных VAM является очень несбалансированной.

Эмоции и их оценки носят субъективный характер. Вот почему важно, чтобы было по крайней мере несколько оценок эмоциональных типов. Даже для человека не всегда очевидно, какую эмоцию он слышит в голосе.

Чтобы уменьшить погрешности данного рода при обучении алгоритма, необходимо, чтобы база данных была сбалансированной и имела большое количество обучающих данных для более точного выделения признаков, соответствующих тому или иному эмоциональному типу.

Заключение

Просмотрев исследования по распознаванию эмоций и результат этих работ, можно сделать следующие выводы. Во-первых, во многом успешность работы алгоритма зависит от качества обучающей базы данных. В ней должны быть представлены все типы эмоций, произнесённые экспертами, причём в равных пропорциях. Во-вторых, большое количество эмоциональных типов резко снижает качество распознавания. Наилучшие результаты дают алгоритмы, выделяющие лишь 2 эмоции. Решение этой проблемы для шести типов эмоций было предложено К. Паком (рисунок).

Таблица 3

Описание баз данных

База данных	Язык	Полная длина (мин)	Количество эмоций	Продолжительность файлов		Продолжительность эмоционального уровня		Комментарии
				средняя (с)	стандартная (с)	средняя (с)	стандартная (с)	
Berlin	Нем.	24,7	7	2,7	1,02	212,4	64,8	Актёры
SAVEE	Англ.	30,7	7	3,8	1,07	263,2	76,3	Актёры
VAM	Нем.	47,8	4	3,02	2,1	717,1	726,3	Не актёры

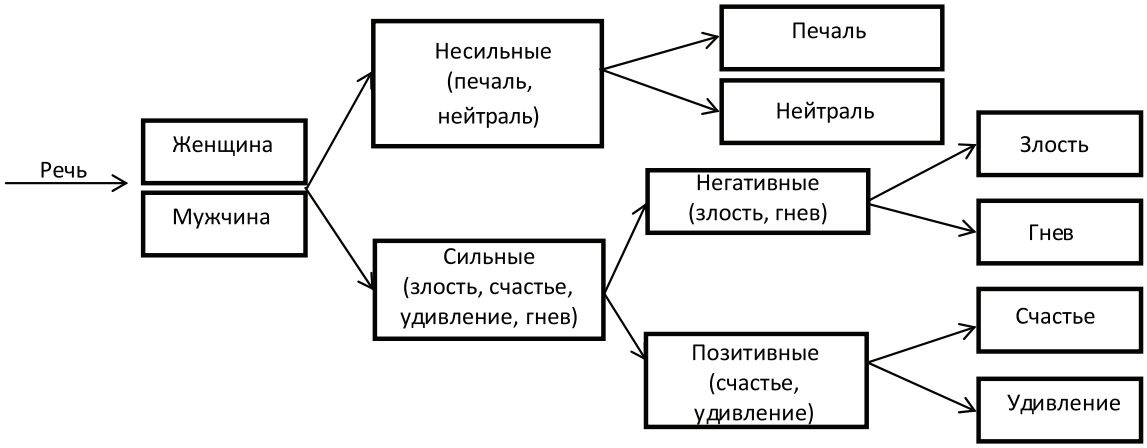

Схема алгоритма классификации эмоций К. Пака

Здесь алгоритм представлен в виде «дерева», в котором поэтапно сравнивается по 2 типа. Это уменьшает вычислительную сложность, что повышает быстродействие и, конечно, увеличивает точность распознавания.

И, в-третьих, на наш взгляд, во многом успех работы зависит от выбора экспертом данных для обучения, так как все эмоциональные типы в этих данных должны быть ярко отличимы друг от друга, должны успешно и достоверно определяться на слух подавляющим большинством экспертов.

На наш взгляд, данная работа поможет в дальнейшем правильно подготавливать данные для проверки работы алгоритма распознавания эмоционального окраса речи, а также, возможно, приведёт к созданию единой и всеобщей эмоциональной базы данных.

Список литературы

1. Алешин Т.С. Новые признаки эмоционально окрашенной речи в речевом сигнале // Научное обозрение. – 2015. – № 1. – С. 102.

2. Alam M.J., Attabi Y., Dumouchel P. Amplitude Modulation Features for Emotion Recognition from Speech // Proc. INRS-EMT. – 2013. Canada.

3. Cernak M.and Wellekens C. Emotional Aspects of Intrinsic Speech Variabilities in Automatic Speech Recognition // Proc. Of 11th Int. Conf. Speech and Computer. – Saint-Petersburg, 2006. – P. 405–408.

4. Dai K., Fell H.J. and MacAuslan J. Recognizing Emotion in Speech Using Neural Networks // In Proceedings of the IASTED International Conference on Telehealth / Assistive Technologies, Ronald Merrell (Ed.). – ACTA Press. 2008. Anaheim, CA, USA. P. 31–36.

5. Metze F., Batliner A., Eyben F. Emotion Recognition using Imperfect Speech Recognition // in Proc. Of the Annual Conference of the International Speech Communication Association (Interspeech 2009). – 2009. Makuhari, Japan. – P. 1–6.

6. Neiberg D., Elenius K., Karlsson I. Emotion Recognition in Spontaneous Speech // In Working Papers 52: Proceedings of Fonetik. – Lund, Sweden, 2006. – P. 101–104.

7. Sidorov M., Brester C., Minker W. Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm // Proceedings of the 9th edition of the Language Resources and Evaluation Conference. – 2014. Reykjavik, Iceland.