LITTERA SCRIPTA MANET
ПОИСК
Еженедельная газета научного сообщества

Инфосфера
№ 18-19 (2017)
05.05.2017

К чему слова? Большую часть информации человек передает молча. Создаются автоматизированные системы, определяющие эмоции человека.

“Важно не что говоришь, а как” - эта простая известная истина отражает суть работы, которой занимается ИТ-специалист и паралингвист, заведующий Лабораторией речевых и многомодальных интерфейсов Санкт-Петербургского института информатики и автоматизации РАН доктор технических наук Алексей КАРПОВ. Со своей научной командой он разрабатывает систему распознавания эмоций и некоторых других состояний человека по речи.

- Наша тема - самая актуальная и востребованная область компьютерной (математической) паралингвистики - новой междисциплинарной отрасли науки, в рамках которой изучаются различные невербальные аспекты в естественной речи, текстах и многомодальной коммуникации, - рассказывает Алексей Анатольевич. - Мы изучаем индивидуальные особенности произношения, акценты, интонации, параметры голоса говорящего (или, как мы говорим, диктора) - в зависимости от пола, возраста, роста, эмоционального состояния и самочувствия. Паралингвистику не следует отождествлять с классической математической лингвистикой, которая ориентирована, в основном, на анализ текстовых данных.

Благодаря различным исследованиям известно, что обычный человек в среднем говорит 10-15 минут в день, это чистая речь без пауз. При этом информация, передаваемая людьми при помощи слов, составляет лишь 5-10% от общего объема данных в процессе межчеловеческой коммуникации. На долю невербальных сигналов, таких как мимика, позы, жесты, касания, запахи, приходится свыше половины всей передаваемой информации, значительна и голосовая паралингвистическая составляющая (жесты, позы, мимика, которые сопровождают речь).

В своей работе мы используем современные методы информационных технологий, искусственного интеллекта и машинного обучения для автоматического распознавания эмоций в речи человека. Базовыми из них ученые считают радость, грусть, злость, страх, отвращение, нейтральное состояние, а также их производные. Точное количество эмоциональных состояний дикторов непостоянно, оно зависит от культуры, образования, окружающего контекста, различных условий.

В отличие от других речевых технологий (например, автоматического распознавания и понимания речи, синтеза речи по тексту,

машинного перевода), системы паралингвистического анализа речи и распознавания эмоций не привязаны к конкретному языку. Поэтому возможно создание практически универсальных методов обработки невербальной акустической информации, конечно, с учетом того, что способы и качество выражения эмоций в некоторой степени отличаются у разных народов и культур.

Все эмоции человека можно представить в двухмерной системе координат, предложенной американским психологом Джеймсом Расселом еще в 1980-х. Одна ось - тон эмоции (valence), которая может быть положительной (приятные чувства) или отрицательной (неприязнь). Вторая ось - интенсивность или сила эмоции (arousal), которая может выражаться сильнее (возбуждение) или слабее (спокойствие).

- Как выглядит автоматическая система?

- Особенность нашего исследования в том, что мы используем для распознавания эмоций человека по его голосу только аудиоинформацию. Сейчас более развиты технологии видеораспознавания эмоций человека по анализу его лица и мимики, уже существуют такие коммерческие системы за рубежом и в России. Разработке же систем аудиораспознавания эмоций уделяется значительно меньше внимания, но они перспективны во многих практических приложениях, например в телефонных контакт-центрах и мобильных диалоговых системах, когда нужно определить текущее психоэмоциональное состояние клиента или оператора: например, доволен ли абонент качеством обслуживания и правильно ли с ним общается оператор в ходе диалога.

Основные компоненты автоматической программной системы, которую мы разрабатываем, - это модули обработки аудиосигналов, вычисления акустических признаков речи, выбора информативных признаков и классификации эмоций с помощью обученных моделей. Мы используем самые современные методы машинного обучения, включая глубокие искусственные нейронные сети (deep neural networks), содержащие несколько скрытых слоев искусственных нейронов.

- Как можно распознавать эмоции человека по речи?

- Мы используем расширенный набор акустических признаков речи, состоящий более чем из 6 тысяч различных компонентов. Это очень большое пространство, которое включает в себя различные энергетические, спектральные и просодические (связанные с ударением) признаки: значения частоты основного тона и формант (резонансные частоты голосового тракта), энергию сегментов речевого сигнала, спектр аудиосигнала, мел-частотные кепстральные коэффициенты (вид спектральной обработки аудиосигнала).

Когда применяются системы автоматического распознавания речи, обычно изучаются короткие речевые фрагменты, такие как фонемы (звуки речи). Акустические признаки, коррелирующие с изменениями психоэмоционального состояния диктора, проявляются на более длительных временных сегментах, поэтому мы вычисляем информативные признаки для целых фраз или слов. При работе с таким пространством возникает проблема больших данных (big data), которую необходимо решать программными и аппаратными способами.

- Насколько точны результаты распознавания? Какова вероятность ошибки?

- Безошибочно распознавать эмоции пока невозможно. Методы обработки сигналов и машинного обучения еще несовершенны. Вариативность выражения психоэмоциональных состояний в речи диктора очень большая, все это крайне сложно учесть при моделировании. Вероятностные модели, которые мы используем, основаны на имеющихся данных, поэтому для совершенствования таких моделей нужны большие речевые корпуса (базы данных), содержащие примеры выражения эмоций многими дикторами.

У нас в лаборатории есть несколько речевых корпусов (баз данных записей речи), содержащих эмоционально окрашенную речь дикторов, которые мы используем в своих исследованиях. Это корпус русской речи RUSLANA, немецкой речи EmoDB, турецкой речи BUEmoDB, английской речи eNTERFACE и несколько других. По результатам предварительных экспериментов на этих данных мы получили точность распознавания эмоций диктора порядка 65-80% для 4-6 базовых эмоциональных состояний.

Некоторые эмоции распознаются лучше (например, радость), другие - хуже (например, страх). Однако в этих базах, в основном, представлены не естественные, а имитированные эмоции, которые были сыграны актерами или обычными людьми в рамках предлагаемого сценария. Такие эмоции распознавать проще, так как иногда они преувеличены. Мы в своих исследованиях хотим сделать акцент на распознавании естественных эмоций в речи человека, которые выражены в реальных ситуациях привычными способами. Но такие данные собирать и обрабатывать намного сложнее.

В нашей работе мы кооперируемся с психофизиологами из группы по изучению детской речи СПбГУ, которая записывает речи и диалоги детей - школьников и дошкольников. Опираясь на данные базы эмоциональной русской речи EmoChildRu, мы смогли определить эмоциональное состояние детей 3-7 лет (нейтральное, комфорт, дискомфорт), а также их возраст. Пока, правда, с вероятностью 55-65%. Этот показатель может показаться невысоким, однако он немного превышает точность, с которой по тем же самым записям распознавались эмоции и возраст детей без применения наших технологий. Так что с этой задачей машина справляется даже лучше человека.

- Есть ли у ваших исследований заказчики?

- Наш проект во многом фундаментальный. Практические результаты, которые получим, мы планируем внедрять “в жизнь” с помощью коммерческих компаний и государственных организаций. Как я уже говорил, они могут пригодиться в автоматизированных телефонных контакт-центрах для анализа речи звонящих абонентов и операторов (в том числе для криминалистики), мобильных приложениях и речевых помощниках для смартфонов, различных интеллектуальных диалоговых и справочных системах (на базе речевых интерфейсов), в полиграфах, социальной робототехнике и других актуальных областях науки и техники. В дальнейшем считаю перспективным создание многомодальной системы распознавания эмоций в результате добавления новых типов информации (модальностей): видеоанализ лица, мимики и поведения человека. Это улучшит точность бесконтактного распознавания психоэмоционального состояния индивида.

- Как уровень вашей работы соотносится с достижениями зарубежных ученых?

- Мы хотели бы достичь результатов мирового уровня и превзойти их. Последние пару лет мы в одной команде с нашими коллегами-друзьями из Турции участвуем в международных соревнованиях (де-факто чемпионат мира) по компьютерной паралингвистике “Computational Paralinguistics ChallengE” ComParE (http://compare.openaudio.eu), которые с 2009 года проводятся в рамках самой престижной конференции по речевым технологиям INTERSPEECH. Эта конференция собирает каждый год почти 2 тысячи ведущих ученых и инженеров со всего мира.

Так вот: два раза наша российско-турецкая команда становилась победителем паралингвистических конкурсов этих соревнований! Мы одержали победу в соревновании ComParE-2015 в рамках INTERSPEECH в Дрездене (Германия), а также в Сан-Франциско (США) в сентябре 2016 года. Так что можно говорить о том, что мы фактически обладаем одними из самых совершенных в мире методов паралингвистического анализа речи и распознавания психоэмоционального состояния дикторов.

К слову, в каждом таком соревновании принимали участие более 30 ведущих научных коллективов со всего мира. В этом году мы снова участвуем в соревнованиях ComParE, итоговые результаты будут оглашены в конце августа на 18-й Международной конференции INTERSPEECH в Стокгольме.

Хотелось бы когда-нибудь эту конференцию провести у нас в России. Наш директор член-корреспондент РАН Рафаэль Мидхатович Юсупов эту идею активно поддерживает, но получить право на ее проведение ничуть не проще, чем получить право на проведение Олимпийских игр. Так что нам придется активно работать.

Василий ЯНЧИЛИН
Иллюстрации предоставлены А. Карповым