СИСТЕМА АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ЭМОЦИЙ ПО РЕЧЕВОМУ СИГНАЛУ

Брановицкий А. А., Борисов Д. В.
Кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: {art7yom, divlboris}@gmail.com

В статье описаны основные области применения и особенности создания систем распознавания эмоций по речевому сигналу. Приведён аналитический обзор актуальных методов для выделения характерных особенностей речевого сигнала. Проиллюстрировано влияние выбора модели нейронной сети на точность распознавания эмоционального состояния.

Введение

На современном этапе развития общества происходит интенсивное развитие информационных технологий, которые вносят качественные улучшения во все сферы человеческой жизни. Одной из перспективных задач в данной области является повышение качества взаимодействия между субъектами в системах взаимодействия человек-компьютер, человек-человек.

Поскольку идентификация эмоционального состояния человека играет важную роль для эффективной коммуникации с ним, задача автоматического распознавания эмоций человека на основе характеристик речевого сигнала представляет интерес не только в теоретическом плане, но и для решения различных прикладных задач.

Распознавание эмоций человека по речи может найти применение в системах взаимодействия человек-компьютер (эмоциональное окрашивание речи операторов автоматизированных колл-центров; улучшение коммуникации людей с голосовыми помощниками, системами виртуальной реальности) или человек-человек (автоматические переводчики, передающие эмоции, заложенные в переводимую речь; детекторы лжи; диагностика психических расстройств на основе изменения эмоционального фона за период времени; мониторинг настроения толпы).

I. Структура системы распознавания эмоций по речевому сигналу

Как и любая типичная система распознавания, система распознавания эмоций на основе речи соотносит исходные данные на входе – речевой сигнал, к определённому классу на выходе – виду эмоции, с помощью выделения существенных признаков – речевых особенностей, характеризующих эти данные.

На рисунке 1 представлена схема работы стандартной системы распознавания эмоций по речевому сигналу. В работе системы можно выделить четыре основных этапа:

1. предварительная обработка сигнала;
2. выделение характерных особенностей речевого сигнала;
3. предобработка особенностей речевого сигнала перед классификацией;
4. классификация.

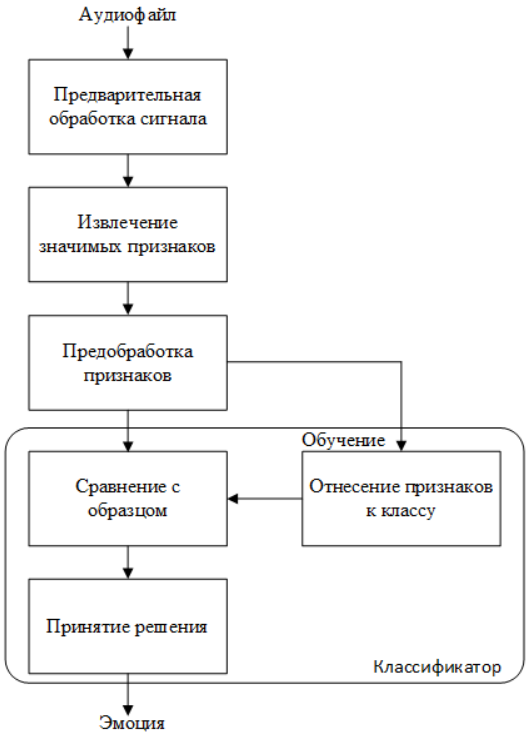

Рис. 1 – Схема работы системы распознавания эмоций

Вход системы – аудиофайлы, которые подвергаются предварительной обработке. Следующим шагом является извлечение значимых признаков из речевого сигнала. Затем применяются алгоритмы фильтрации, отсеивающие малозначимые признаки. Далее вектор признаков поступает на вход классификатору. Выход системы – вид эмоции.

II. Методы, применяемые для предварительной обработки сигнала

Предварительная обработка сигнала является важным этапом системы, так как оказывает огромное влияние на производительность классификатора. Очень важно подавать на вход нейронной сети данные, отражающие ключевые

особенности образца для классификации. Предварительная обработка охватывает аналогово-цифровое преобразование (АЦП) входного сигнала для получения его спектральных составляющих; цифровую фильтрацию сигнала, нормализацию, сегментацию сигнала, удаление неречевых компонентов сигнала [1].

Оцифрованные образцы речи сначала нормализуются. Далее образцы сегментируется на кадры продолжительностью 30 мс с перекрытием 10 мс с использованием окна Хэмминга. Наконец, удаляются кадры с тишиной и невокализованные кадры.

III. Характерные особенности речевого сигнала и методы их расчёта

Для проведения успешной классификации чрезвычайно важно извлечь значимые признаки из речевого сигнала. Наиболее эффективные особенности речевого сигнала для классификации на данный момент времени сводятся к следующему списку:

- Частота основного тона (англ. pitch) – это частота колебания голосовых связок при произнесении тоновых звуков.
- Форманты (англ. formants) – группа усиленных обертонов, формирующих специфический тембр голоса. Рассчитываются при помощи метода линейного прогнозирующего кодирования (англ. Linear Predictive Coding, LPC).
- Мел-кепстральные коэффициенты (англ. Mel-frequency Cepstral Coefficients, MFCCs). Мел – единица высоты звука, основанная на восприятии этого звука нашими органами слуха [2].
- Скорость пересечения нуля (англ. zero crossing rates, ZCR) – это скорость изменения знака вдоль сигнала, то есть скорость, с которой значение сигнала изменяется с положительного на отрицательное или с отрицательного на положительное.
- Джиттер и шиммер (англ. jitter and shimmer). Джиттер – мера пертурбации (возмущений) частоты основного тона, показывающая непроизвольные изменения в частоте смежных вибрационных циклов голосовых складок. Шиммер – мера аналогичная джиттеру, только характеризующая пертурбации амплитуд сигнала на смежных циклах колебаний основного тона.
- Энергия сигнала (англ. energy). Рассчитывать необходимо краткосрочный энергетический контур, так как он напрямую связан с уровнем эмоционального возбуждения.
- Длительность и вокализация (англ. duration and voicing). Вокализация – отношение количества вокализованных к количеству невокализованных кадров.

IV. Предобработка признаков речевого сигнала

Векторы признаков, полученные на предыдущем этапе, перед подачей их на вход выбранному классификатору нормализуются; содержащие более 2–10% нулевых значений, отбрасываются. Так как на большинство классификаторов негативно влияет избыточность, то для уменьшения размерности итогового вектора признаков используется алгоритм выбора признаков (прямой выбор признаков, генетический алгоритм, последовательный прямой плавающий поиск).

V. Классификация

Классификация – заключительный этап работы системы распознавания эмоций. Точность классификации в значительной мере зависит от выбранного типа классификатора. Ранее в качестве классификатора для систем распознавания эмоций применялись Скрытые Марковские модели (англ. HMM), метод k-ближайших соседей (англ. k-NN), Модель гауссовых смесей (англ. GMM), метод опорных векторов (англ. SVM), но в данный момент отдаётся предпочтение нейронным сетям различных архитектур [3].

VI. Оценка эффективности системы

При практической реализации системы для сокращения размерности вектора признаков применялся алгоритм прямого выбора признаков (англ. Forward Feature Selection, FFS), в качестве классификаторов выступали нейронные сети. Таблица 1 демонстрирует влияние типа и архитектуры нейронной сети на итоговую эффективность разрабатываемой системы.

Таблица 1 – Успешность классификации эмоций нейронными сетями

ANN	Ang	Bore	Fear	Hap	Sad	Avg
FFBP	74%	67%	77%	84%	69%	74%
FIT	71%	75%	62%	68%	70%	69%
CF	69%	72%	70%	70%	64%	69%
LVQ	58%	78%	70%	63%	80%	70%
PNN	96%	83%	76%	88%	79%	84%

В целом, выбор подходящей нейронной сети позволяет создать систему распознавания эмоций по речевому сигналу, обладающую довольно высокой точностью.

Список литературы

1. Rong, J. Acoustic Features Extraction for Emotion Recognition / J. Rong // 6th IEEE/ACIS International Conference. – 2007. – P. 419–424.
2. Patil, K. J. Emotion Detection From Speech Using Mfcc & Gmm / K. J. Patil, P. H. Zope, S. R. Suralkar // IJERT. – 2012. – Vol. 1, № 9. – P. 101–103.
3. Hamidi, M. Emotion Recognition From Persian Speech With Neural Network / M. Hamidi, M. Mansoorizade // IJAIA. – 2012. – Vol. 3, № 5. – P. 107.