Технология распознавания эмоций человека по речи может быть использована во многих интересных областях. Не удивительно, что существует множество исследований на эту тему. В этой статье рассматриваются исследования по распознаванию эмоций в речи. Выделяются отличия некоторых подходов и раскрываются плюсы и минусы этих подходов. Затем, на основе проанализированной информации предлагается реализация упрощенной системы распознавания базовых эмоций

1 Введение

Люди выражают эмоции не только с помощью жестов и мимики. Эмоции также выражаются в речи. Когда человек, например, возбужден, тон голоса становиться выше, растет темп речи и средняя длительность слогов уменьшается. Такого рода характеристики, связанные с эмоциями описаны во многих исследованиях в этой области.

Мы способны распознавать базовые эмоции независимо от языка, возраста и пола говорящего. Это означает, что существует связь между динамикой звуковых характеристик и выражаемыми эмоциями. И это доказывает множество работ. Значит, возможно реализовать такую технологию распознавания эмоций, которая работает независимо от языка, возраста и пола говорящего.

Почему такого рода технология может быть полезна? На самом деле, данное решение может быть использовано во многих областях. Наиболее из очевидных из которых: при построении роботов и для оценки качества обслуживания клиентов.

Для роботов, которые работают в социальной среде, может оказаться важной и необходимой такая способность. Это относиться к теме эмоциональных вычислений.

Эмоциональные вычисления (affective computing) это изучение и развитие систем и устройств, которые способны распознавать, интерпретировать и имитировать человеческие эмоции. Этот подход к разработке роботов выделяет важность эмоций и доказывается важностью эмоций для людей. Потому что эмоции помогают лучше взаимодействовать людям, эмоциональная окраска позволяет улучшить запоминание, также это помогает при принятии решений.

Во-первых, технология распознавания эмоций может улучшить систему принятия решений. Так как могут возникать ситуации, когда роботу для принятия решения необходимо учитывать эмоциональное состояние человека. К тому же, выявление связей между звуковыми характеристиками и эмоциями, дает возможность реализовать эмоциональную речь робота. Если у говорящего робота будет способность эмоционально окрашивать некоторые выражения, это позволит ему лучше взаимодействовать с людьми. В целом, можно сделать вывод, что робот с возможностью распознавания и имитации эмоций более дружелюбен и эффективен. Уже существуют такие роботы, например, робот Pepper:

Pepper это первый робот, спроектированный для жизни с людьми. К сожалению, он не убирается, не готовит и не имеет супер возможностей. Pepper – социальный робот, способный поговорить с Вами, распознавать и реагировать на ваши эмоции, способный передвигаться и существовать автономно.[2]

Более приближенное к настоящему применение этой технологии лежит в области оценки качества обслуживания клиентов. Например, call-центры ведут записи звонков, и для оценки качества работы операторов, можно анализировать звонки и выявлять особо эмоциональные разговоры. Эта технология может быть использована вместе с распознаванием речи. И для распознавания эмоций говорящего, можно анализировать звуковые характеристики и произнесенные фразы.

2. Цель работы

Существует много работ рассматривающих распознавание эмоций по речи. Их основные отличия в использовании различных: акустических характеристик, признаков, методов классификации, распознаваемых эмоций.

Также существуют статьи, в которых рассматриваются эти отличия, но зачастую в них рассматриваются только акустические характеристики, признаки и методы классификации. Не часто рассматривается, например, влияние на точность распознаваемые эмоции.

Таким образом, главная цель этой работы рассмотреть отличия различных подходов и разработать упрощенный прототип распознавания эмоций по речи. Затем, проверить влияние различных параметров на точность распознавания. Также, рассмотреть способы решения таких проблем как: извлечение звуковых характеристик, удаление шума, разбиение речи на фразы. Будут рассмотрены конкретные библиотеки и алгоритмы для решения этих проблем.

3. Обзор

3.1

Голос, как и любой другой звук может быть описан некоторыми акустическими характеристиками. Перед тем рассматривать конкретные исследования, стоит рассмотреть эти акустические характеристики:

Тон или фундаментальная частота

Громкость

Форманта

Мел-кепстральные коэффициенты

Тон

Частота человеческого голоса находиться в области между 300 и 3400 Гц. А фундаментальная (основная) частота для мужского пола между 85 и 185 Гц, а для женского пола между 165 и 255 Гц.

Основная частота- частота повторения для периодической функции, определяется (нестрого) как низшая частота сложной периодической волны, иногда называется первой гармоникой

Громкость

Эта характеристика звука также может быть полезна для распознавания эмоций по речи. Но стоит отметить, что необходимо работать с ней с осторожностью, ведь эта характеристика очень сильно зависит от качества записи, от специфичной манеры говорить и других факторов, которые могут способствовать получению ложных данных.

Гро́мкость зву́ка — субъективное [восприятие](https://ru.wikipedia.org/wiki/%D0%92%D0%BE%D1%81%D0%BF%D1%80%D0%B8%D1%8F%D1%82%D0%B8%D0%B5) силы [звука](https://ru.wikipedia.org/wiki/%D0%97%D0%B2%D1%83%D0%BA) (абсолютная величина слухового ощущения). Громкость главным образом зависит от [звукового давления](https://ru.wikipedia.org/wiki/%D0%97%D0%B2%D1%83%D0%BA%D0%BE%D0%B2%D0%BE%D0%B5_%D0%B4%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5) и[частоты](https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%B0) звуковых колебаний. Также на громкость звука влияют его спектральный состав, локализация в пространстве, тембр, длительность воздействия звуковых колебаний и другие факторы

Форманта и мел-кепстральнаые коэфициенты

Часто для распознавания эмоций достаточно тона и громкости голоса. Но для более точных результатов можно использовать и другие характеристики звука. Например, форманты или мел-кепстральные коэффициенты.

**Форма́нта** — термин [фонетики](https://ru.wikipedia.org/wiki/%D0%A4%D0%BE%D0%BD%D0%B5%D1%82%D0%B8%D0%BA%D0%B0), обозначающий [акустическую](https://ru.wikipedia.org/wiki/%D0%90%D0%BA%D1%83%D1%81%D1%82%D0%B8%D0%BA%D0%B0) характеристику [звуков речи](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D1%87%D0%B5%D0%B2%D0%BE%D0%B9_%D0%B7%D0%B2%D1%83%D0%BA) (прежде всего гласных), связанную с уровнем частоты голосового тона и образующую [тембр](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BC%D0%B1%D1%80) звука

3.2 Признаки

Для классификации эмоций используя акустические характеристики речи, необходимо сначала выделить признаки и закономерности этих среди этих характеристик. Таким образом, из обучающей выборки мы вычисляем признаки для каждой эмоции. И когда уже классификатор обучен, классификация проходит следующим образом:

1) извлечь признаки из обрабатываемой записи

2) сравнить эти признаки с признаками определенных эмоций

По тому, как и для кого рассчитываются эти признаки, их можно разделить на две группы:

-зависимые от говорящего

- независимые от говорящего

3.2.1

Зависимые от говорящего признаки рассчитываются для каждого человека отдельно. Поэтому требуется разделенная обучающая выборка. И при классификации необходимы идентификация человека и рассчитанные признаки для его речи. Во время классификации сравниваются полученные признаки с его персональными признаками. Данный подход может оказаться более точным, потому что в отличие от другого подхода, здесь нет такого большого разброса значений. Но данный подход не универсален, поскольку необходимо для каждого человека хранить статистические данные (признаки). Эти признаки основаны на некоторых средних значениях, например, средняя высота тона, средняя громкость, средний диапазон высоты тона и т.д., а эти значения разные для всех людей. Поэтому такие признаки могут быть применены в очень редких случаях.

3.2.2

В отличие от предыдущего подхода разделенная обучающая выборка не требуется. Но поскольку точность классификации сильно зависит от обучающей выборки, она должна содержать как можно больше записей с как можно разными людьми (пол, возраст, родной язык). Эти признаки основаны на динамике акустических характеристик и некоторых средних значениях.

3.3 Классификация

С одной стороны, проблема классификации в этом случае очень похожа на другие случаи и может быть решена с помощью таких классификаторов как: Бинарная диаграмма решений, Наивный Байесовский классификатор, Нейронные сети, Векторная Машина.

Но с другой стороны, присутствуют некоторые специфичные моменты, вызванные похожестью некоторых эмоций и неоднозначностью между ними. Большинство методов классификации очень подробно рассмотрены в статье. Давайте рассмотрим некоторые из них.

3.3.1 Бинарная диаграмма решений

Бинарное дерево принятия решения - это структура данных, которая используется для представления булевой функции. Булева функция может быть представлена как направленный ацикличный граф, состоящий из внутренних узлов решений, каждый из которых имеет по два потомка, и терминальных узлов, каждый из которых соответствует одному из значений булевой функции.

Достоинства: легкая реализация, четка видна связь между входными и выходными данными. Обучение и классификация простые и быстрые. Точность сравнима с другими классификаторами на простых данных. Возможность преобразовать в простые правила классификации.

Недостатки: при обучении могут получиться слишком сложные деревья, которые не обобщают реальные данные. Проблема построения оптимального дерева – NP полная проблема. Алгоритмы обучения построены на эвристических алгоритмах таких как жадные алгоритмы. Где на каждом шаге принимается локальное оптимальное решение. Такие алгоритмы не обязательно приводят к глобально оптимальным решениям.

3.3.2 Искусственные нейронные сети

Искусственные нейронные сети – [математическая модель](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8C), а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования [биологических нейронных сетей](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D1%81%D0%B5%D1%82%D1%8C) — сетей [нервных клеток](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD) живого организма

3.4 Распознаваемые эмоции

Точность распознавания эмоций зависит не только от выделенных признаков и методов классификации, это также зависит от распознаваемых эмоций. Это вызвано тем, что некоторые эмоции выражаются очень похожим образом. Чем больше будет выбрано эмоций для распознавания, тем, скорее всего, будет меньше точность их распознавания. Поэтому это некий компромисс между количеством распознаваемых эмоций и точностью.

Например, на рис. 1 показана связь между эмоциями и некоторыми признаками акустических характеристик. Видно, что некоторые эмоции очень похожи по некоторым показателям (злость-радость, грусть-скукота). Хотя и количество эмоций не такое большое, уже можно видеть некоторые пересечения в признаках. С увеличением количества эмоций, таких пересечений станет только больше. Поэтому важно определять набор распознаваемых эмоций для каждой решаемой задачи отдельно.

Большинство работ в этой области все больше и больше используют признаки, независящие от говорящего. Выделяя из речи признаки, основанные на динамике речевого сигнала. Поскольку такой подход более универсален и может быть применен во многих отраслях.

Проблема классификации часто решается с использованием: векторных машин, алгоритма классификации ближайшего соседа, скрытых Марковских моделей и нейронных сетей. Хотя более простые методы, такие как: бинарные диаграммы решения, нечеткие множества тоже работают и обладают сравнимой точностью и эффективностью.