**Abstact**

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 трех базовых эмоций: злость, радость и грусть.

**Введение**

Технология распознавания эмоций человека может быть полезна в разных областях. Наиболее очевидные из которых: построение роботов и оценка качества обслуживания клиентов. Если у говорящего робота будет способность эмоционально окрашивать некоторые выражения, это позволит ему лучше взаимодействовать с людьми. Также, чтобы робот мог хорошо взаимодействовать с людьми, ему необходимо хорошо понимать людей. Могут возникать ситуации, когда при принятии определенных решений, робот должен учитывать эмоциональное состояние человека. Например,(ПРИВЕСТИ ПРИМЕР).

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.

Существует множество исследований, раскрывающих зависимости между акустическими характеристиками и эмоциями человека. Отличаются эти исследования по различным критериям:

* *Акустические характеристики речи*
* *Признаки или свойства акустических характеристик для различных эмоций*
* *Методы классификации*
* *Распознаваемые эмоции*

Главной задачей этой работы является выяснить как влияют эти критерии на точность распознавания и предложить свою реализацию.

**2. Обзор исследований**

*2.1 Акустические характеристики*

Прежде чем приступать к рассмотрению конкретных исследований, стоит рассмотреть акустические характеристики речи:

* *Основная частота*
* *Громкость*
* *Форманты*
* *Мел-кепстральные коэффициенты (MFCC)*

*Основная частота (fundamental frequency)* – или по-другому высота тона речи. Человеческая речь по частоте примерно расположена в диапазоне от 300 до 3400 Гц. А *основная частота* речи у мужчин расположена от 85 до 185 Гц, у женщин от 165 до 255.

Наиболее часто в исследованиях используют только *высоту тона* и *громкость.* Хотя такие характеристики как *форманты* и *мел-кепстральные коэффициенты* тоже несут полезную информацию.

*Форманта- термин фонетики, обозначающий акустическую характеристику звуков речи (прежде всего гласных), связанную с уровнем частоты голосового тона и образующую тембр звука (https://ru.wikipedia.org/wiki/Форманта).*

*Мел-кепстральные коэффициенты(MFCC)- часто используются для характеристики речевых сигналов. Чаще используются для распознавания речи.*

*2.3 Признаки или свойства акустических характеристик*

Для классификации эмоций по акустическим характеристикам необходимо выделить некоторые признаки для определенных эмоций. И эти эталонные признаки рассчитываются на основе обучающей выборки. Таким образом, при классификации речи происходит следующее:

1. Рассчитываются признаки для заданной речи
2. Признаки заданной речи сравниваются с эталонными признаками

По тому как рассчитываются эталонные признаки, их можно разбить на две группы: speaker dependent и speaker independent.

*Speaker dependent*

В этом случае нам необходима обучающая выборка для каждого человека, речь которого будет оцениваться. Этот подход не универсален, и может быть применен в узкой сфере. Хотя и может оказаться, что этот подход может оказаться более точным. Например: средняя частота(тон), средняя громкость, диапазон тона и громкости. При классификации речи нам необходимо знать, чья это речь, и необходимы рассчитанные признаки для этого человека.

*Speaker independent*

Это универсальный подход, потому что все признаки рассчитаны для различных людей. И для точной работы необходима большая обучающая выборка с большим количеством людей. Например, Pitch DDS, Formant DDS, длительность фраз, длительность пауз, DDS (разница-расстояние-наклон) используется для оценки динамики речевого сигнала. Эти признаки более подробно описываются в части:????????.

**2.4. Методы классификации**

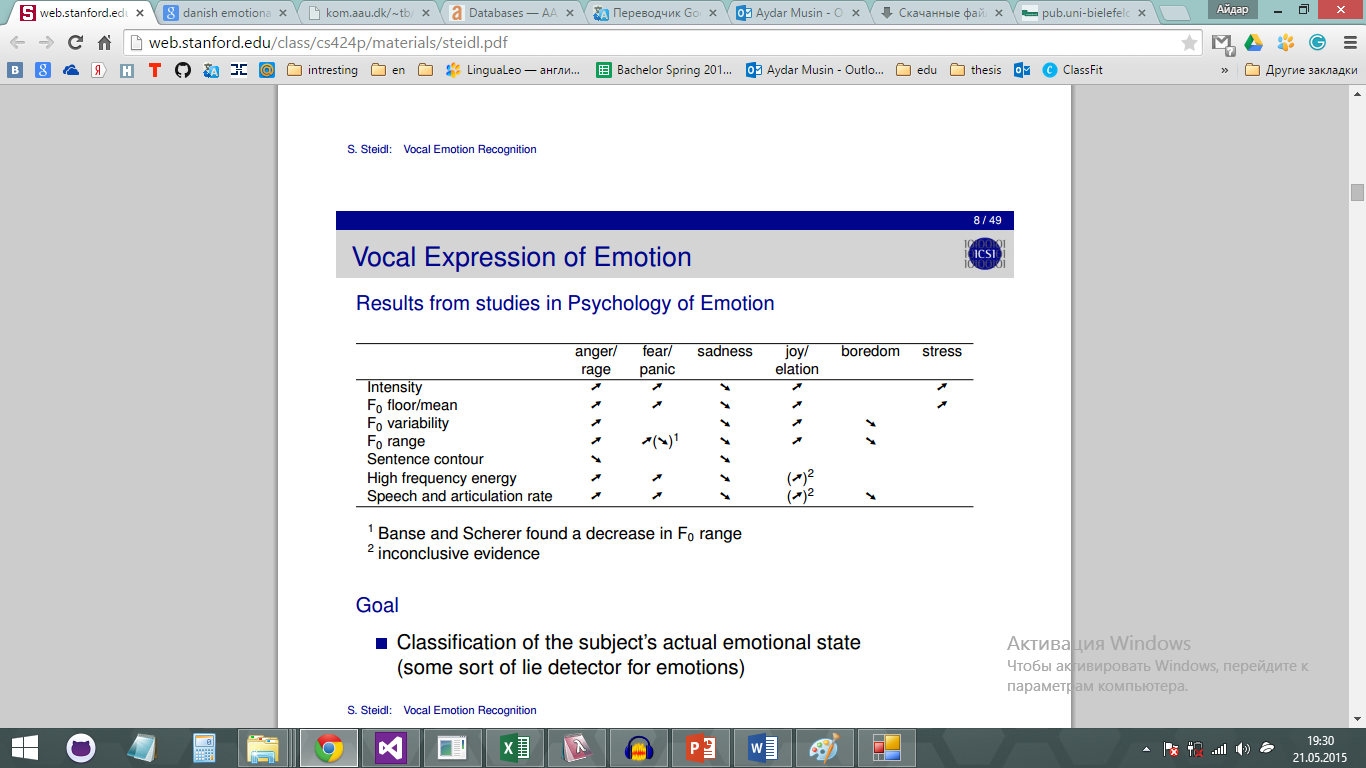
Методы классификации очень хорошо описаны в статье: [Speaker Emotion Recognition Based on Speech Features and Classification Techniques], приведу некоторые из методов в таблице:

*Бинарные деревья принятия решений*

|  |  |  |  |
| --- | --- | --- | --- |
| **Название** | **Описание** | **Достоинства** | **Недостатки** |
| **Binary Decision Tree** | Decision Tree is a stream design drawing like structure in which inward hub talks to check on a quality, every extension talks to deduction of check and every leaf hub talks to class title (choice taken in the wake of registering all traits). A way from root to leaf speaks to arrangement runs the display. In alternative examination a conclusion tree and the almost identified influence journal is utilized as a visual and scientific choice support device, where the usual qualities (or required utility) of arguing options are computed. | Easy implementation, easy explanation of input and output relationship Can handle high dimensional data Easy to interpret for small sized trees The learning and classification steps of induction are simple and fast Accuracy is comparable to other classification techniques for many simple data sets Convertible to simple and easy to understand classification rules | Decision-tree learners can create overcomplex trees that do not generalize the facts and figures well. Decision trees can be unstable because small variations in the facts and figures might outcome in a absolutely different tree being developed. This difficulty is mitigated by using decision trees inside an ensemble. The difficulty of discovering an optimal decision tree is known to be NPcomplete under several facets of optimality and even for easy concepts. Consequently, functional decision-tree learning algorithms are founded on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms will not assurance to return the globally optimal decision tree. There are concepts that are hard to discover because decision trees do not articulate them effortlessly, such as XOR, parity or multiplexer troubles. conclusion tree learners conceive biased trees if some categories dominate. It is thus suggested to balance the dataset prior to fitting with the conclusion tree |
| **Artificial Neural Network** | An Artificial Neural mesh (ANN) is a facts and figures organising standard that is inspired by the way biotic anxious structures, for example the cerebrum, process facts and figures. The key constituent of this ideal model is the innovative structure of the facts and figures handling structure. It is made out of countless interconnected changing components (neurones) employed as one to tackle specific issues. ANNs, for demonstration persons, study by illustration. An ANN is designed for a specific provision, for demonstration design acknowledgement or information characterization, through a revising method. revising in living structures includes acclimations to the synaptic associations that exist between the neurones. This is accurate of ANNs besides | They can both about any convoluted conclusion supplied that enough nodes are utilised. Neural systems are rather easy to implement (you do not need a good linear algebra solver as for examples for SVNs). Neural networks often exhibit patterns alike to those exhibited by humans. although this is more of interest in cognitive sciences than for functional examples | Long preparing time The VC measurement of neural systems is indistinct. This is extremely critical when you need to think about how exceptional an answer could be. Neural systems can't be retrained. Provided that you include information later, this is just about difficult to add to an existing system. Taking care of time arrangement information in neural systems is an exceptionally confounded point. |
| **K- Nearest Neighbor** | All occurrences contrast to focuses in a n-dimensional Euclidean space. Alignment is postponed till another occasion reaches. Alignment finished by investigating emphasizes vectors of the characteristic focuses. goal capacity may be discrete or authentic esteemed | No training is required, confidence level can be obtained | Classification correctness is reduced is convoluted decision-region boundary lives, sizable storage needed KNN algorithm is that it is a slovenly learner, i.e. it does not discover any thing from the teaching data and easily values the teaching data itself for classification The algorithm should compute the expanse and sort all the teaching data at each proposition, which can be slow if there are a large number of teaching demonstrations The algorithm does not learn anything from the teaching facts and figures, which can outcome in the algorithm not generalizing well and furthermore not being robust to loud data |
| **Naïve Bayes classifier** | The Naive Bayes Classifier procedure is reliant upon the purported Bayesian hypothesis and is particularly suited when the dimensionality of the inputs is high. Notwithstanding its effortlessness, Naive Bayes can frequently outflank more refined grouping strategies | Fast to train (single scan). Fast to classify Not sensitive to irrelevant features Handles real and discrete data Handles streaming data well | Assumes independence of features |

**2.5 Распознаваемые эмоции**

Точность распознавания зависит не только от выделяемых признаков и методов классификации. Но и зависит от выбранных распознаваемых эмоций. Вызвано это тем, что некоторые эмоции очень похожи. И чтобы правильно классифицировать похожие эмоции необходимо либо вводить новые признаки, либо улучшать методы классификации. Поэтому это компромисс между количеством распознаваемых эмоций и точностью. И необходимое количество эмоций зависит от требований конкретной задачи.



Например, эта таблица из статьи [Stefan Steidl; October 26, 2010;” Vocal Emotion Recognition”; ICSI].

На этой таблице видно, что по многим признаками такие эмоции как злость(anger) и радость(joy) очень похожи. Поэтому довольно важно определить набор эмоций, которые необходимо распознавать для конкретной задачи.

**Примеры конкретных исследований**

**Реализация**

**Постановка задачи**

Главной задачей является разработать упрощенную систему распознавания эмоций по речи. Система должна позволять определять эмоции человека по речи записанной на звуковой файл или с микрофона. При этом необходимо, чтобы программа была устойчива к малым помехам или шуму. Реализация будет осуществляться на языке программирования C#.

Задачу реализации можно разбить на части:

1. Считывание акустических характеристик с звукового файла или микрофона
2. Разбиение речи на фразы или отдельные слова
3. Получение определенных признаков акустических характеристик
4. Классификация

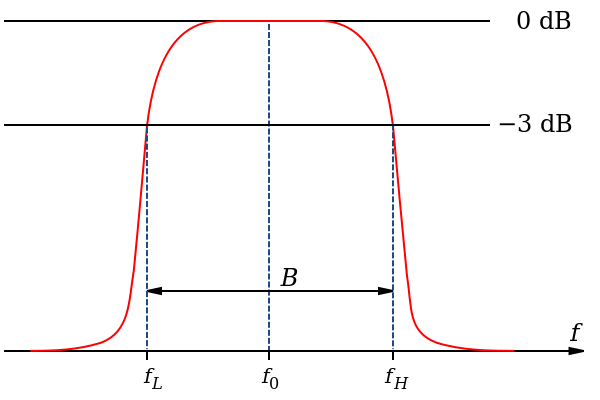
*1. Получение акустических характеристик*

Начнем с того, что для наших целей необходимы характеристики только голоса. Поэтому, чтобы избежать ложных результатов, сначала нужно исключить лишние шумы. Это важно не только для получения правильных характеристик голоса, это важно еще для того, чтобы правильно выделить промежутки молчания. Потому что это позволит делить речь на фразы и рассчитывать темп речи.

Чтобы получить данные с звуковых файлов или микрофона можно использовать библиотеки:

* NAudio (http://naudio.codeplex.com)
* Bass Audio library (http://naudio.codeplex.com)

Но эти библиотеки не позволяют получить акустические характеристики напрямую. Поэтому придется еще реализовывать расчет характеристик. Например, для расчета высоты тона придется реализовывать БПФ (быстрое преобразование Фурье). К тому же, придется реализовывать функции фильтрации звука от шума. Самую простую фильтрацию шума можно сделать с помощью pass-band filter



[fL – fH] –это диапазон необходимых частот. Частоты, выходящие за рамки будут игнорироваться.

Но для наших целей больше подходит пакет PRAAT (praat.org). Этот пакет часто используется для работы с речью и содержит все необходимые функции для нашего проекта:

* Получение основной частоты
* Получение громкости
* Получение формант
* Получение спектра
* Функции фильтрации

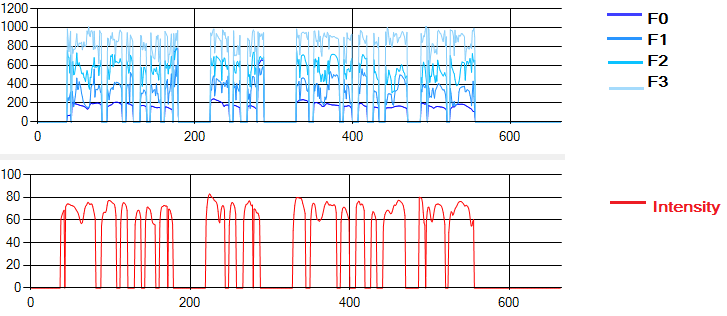
Для использования всех функций пакет имеет свой скриптовый язык. Работу с скриптами подробно рассматривать не будем, поскольку эту информацию можно найти на сайте praat.org.

Для удаления шумов в библиотеке есть функция remove noise[???], которая включает в себя pass band фильтр и специальные функции уменьшения шума.

С помощью пакета PRAAT мы будем получать такие характеристики как:

* Основной тон
* Громкость
* 1,2,3 форманты
* Центр тяжести для спектрограммы

На выходе получаем характеристики для каждого момента времени:



Где, F0-основной тон, F1 первая форманта, F2-вторая форманта, F3-третья форманта, Intensity- громкость.

На графиках видно, что есть участки, где значения всех характеристик равны нулю. Так отмечаются участки молчания. Т.е. когда частота звука выходит за пределы человеческого голоса, фильтр присваивает значение ноль для характеристики основного тона, соответственно для остальных характеристик значения тоже обнуляются.

*2. Разбиение речи на фразы или слова*

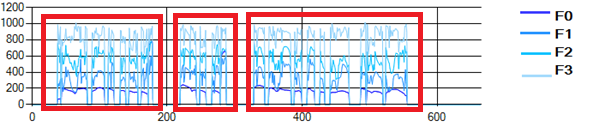
Поскольку человек при разговоре может придавать эмоциональную окраску отдельным фразам и словам, нам необходимо разделить речь на фразы и рассчитывать признаки для них отдельно.

Стоит отметить, что точное разделение на фразы невозможно без предварительной фильтрации речи. Поэтому успешность этого этапа зависит от предыдущего.

На выходе их предыдущего этапа мы получаем ряд характеристик, среди которых участки молчания отмечены нулями. Поэтому алгоритм разделения может быть таким:

1. Пока не встретился нуль, собираем значения фразы
2. Как только встретился нуль, закрываем эту фразу и создаем новую и переходим к шагу 1.

Но проблема в том, что нули могут встретится и посередине некоторых фраз или слов. Поэтому необходимо определить минимальные промежутки между фразами. Т.е. какой минимальной длины промежуток должен быть между фразами. Как показано на графике, речь поделена на фразы, но внутри фраз есть промежутки молчания, но они незначительные



Таким образом, в результате получаем ряд характеристик и указатели на начало и конец фраз. Теперь из этих характеристик необходимо выделить признаки для последующей классификации.

*3. Получение признаков акустических характеристик*

Используемые признаки:

|  |  |
| --- | --- |
| Высота тона (F0) | Вариативность |
| Диапазон |
| DDS |
| Форманта 1 (F1) | DDS |
| Форманта 2 (F2) | DDS |
| Форманта 2 (F3) | DDS |
| Громкость | Вариативность |
| диапазон |
| DDS |
| Время | Средняя продолжительность фраз |
| Средняя продолжительность интервалов молчания |
| Спектрограмма | Центр тяжести спектрограммы (centroid) |

*Вариативность*

Вариативность вычисляется через стандартное отклонение ряда.

\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^n\left(x_i-\bar{x}\right)^2}.

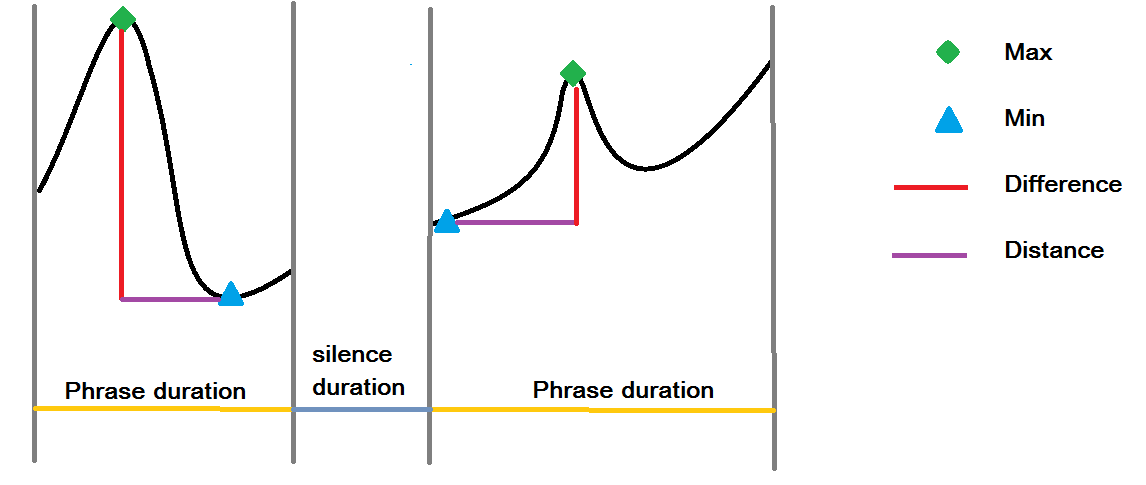
Где \bar{x}\,\! - среднее арифметическое ряда.

*Диапазон*

Диапазон равен разнице между максимальной и минимальной величиной.

*DDS*

Для подсчета DDS в каждой фразе находим локальный максимум и минимум. Берем разницу между максимумом и минимумом, также берем дистанцию между ними. Эти две характеристики тоже описывают вариативность сигнала и его динамику. Например, когда график затухает, дистанция между максимумом и минимумом будет отрицательной.



Продолжительности фраз и интервалов молчания вычисляются по указателям начала и конца фраз. А центр тяжести спектрограммы сразу вычисляется пакетом PRAAT.

*Получение реальных признаков на основе обучающей выборки*

В качестве обучающей выборки будем использовать базу эмоциональных записей EmoDB (<http://emodb.bilderbar.info>). Мы будем работать с тремя базовыми эмоциями: радость, злость, грусть, нейтральность.

Результаты получения признаков:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | anger | happy | sadness | neutral |
| PitchDIf | 137,08 | 211,55 | 67,32 | 64,63 |
| PitchDis | -6 | -10,125 | -19,33 | -8 |
| IntDif | 22,22 | 22,92 | 19,63 | 18,76 |
| IntDis | 6,44 | -5,83 | -7,41 | -3,2 |
| F1Dif | 239,44 | 228,16 | 210,52 | 193,49 |
| F1Dis | -1,71 | -5,87 | -9 | -1,66 |
| F2dif | 349,82 | 318,56 | 340,58 | 299,02 |
| F2Dis | -1,42 | -3,15 | -2,32 | 0,66 |
| F3Dif | 355,51 | 359,38 | 449,35 | 425,46 |
| F3Dis | -0,4 | -2,28 | -0,33 | 2,6 |
| PitchRange | 465,16 | 406,37 | 150,065 | 457,2 |
| IntRange | 31,36 | 28,95 | 28,865 | 27,61 |
| PitchVariance | 7253,96 | 10823,92 | 1737,46 | 6720,77 |
| IntVariance | 46,19 | 41,20 | 35,10 | 36,24 |
| PhraseDuration | 28,9 | 38,16 | 39,58 | 24,66 |
| SilenceDuration | 5,2 | 4,45 | 5,375 | 5,4 |
| Centroid | 580,3 | 519,54 | 359,08 | 350,02 |

Сокращения в таблице: Dif-Difference –разница, Dis-Distance-дистанция, Int-Intensity-Громкость

Как видно по таблице, эмоции злость и радость очень близки по многим показателям. Тоже самое с эмоциями грусть и нейтральность. Хотя есть признаки, которые показывают разницу.

*4.Классификация*

Мы получили некие значения признаков для определенных эмоций, теперь необходимо построить классификатор. Мы не можем выставить четкие рамки для признаков, потому что эти значения получены путем усреднения. И среди них есть значения намного отличающиеся от этого значения. В тоже время некоторые признаки расположены очень в плотную и пересекаются, и нельзя просто разделить значения этого признака на 4 класса.

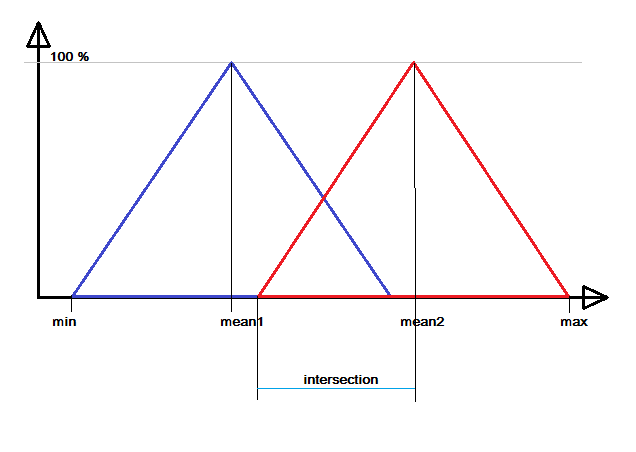
Поэтому было решено использовать нечеткие множества для классификации каждого признака по классам: «низко» и «высоко».

*Построение множеств для классификации*

Для каждого признака мы получаем 4 значения, соответствующие 4 эмоциям. Все эмоции попарно похожи. Например, как говорилось выше, злость-радость и грусть-нейтральность. Поэтому среди 4 значений признака как правило есть 2 «низких» значения и 2 «высоких». Единственное, для каждого параметра нам необходимо знать минимальное и максимальное значение.

Алгоритм построения двух множеств:

1. Сортируем 4 значения
2. mean1=среднее двух первых значений, mean2=среднее двух последних
3. intersection= 20% от максимального элемента



Таким образом строим нечеткие множества для всех характеристик и для каждого признака необходимы две функции IsLow и IsHigh, которые будут возвращать вероятность принадлежности ко множеству. И отталкиваясь от такой модели каждую эмоцию можно охарактеризовать в таком виде:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | anger | happy | sadness | neutral |
| PitchDIf | высоко | высоко | низко | низко |
| PitchDis | низко | высоко | низко | высоко |
| IntDif | высоко | высоко | низко | низко |
| IntDis | высоко | низко | низко | высоко |
| F1Dif | высоко | высоко | низко | низко |
| F1Dis | высоко | низко | низко | высоко |
| F2dif | высоко | низко | высоко | низко |
| F2Dis | высоко | низко | низко | высоко |
| F3Dif | низко | низко | высоко | высоко |
| F3Dis | низко | низко | высоко | высоко |
| PitchRange | высоко | низко | низко | высоко |
| IntRange | высоко | высоко | низко | низко |
| PitchVariance | высоко | высоко | низко | низко |
| IntVariance | высоко | высоко | низко | низко |
| PhraseDuration | низко | высоко | высоко | низко |
| Centroid | высоко | высоко | низко | низко |

Расчет вероятности эмоции выглядеть как:

Anger=PitchDif.IsHigh()+PitchDis.IsLow()+IntDif.IsHigh()+…

Система в целом

