# Recognition of emotions in speech: overview and implementation

Aydar Musin
Supervisor: prof. Maxim Talanov

June 2015, Innopolis University

**Abstract**

In this paper will be considered research on recognition of emotions in speech, distinguished their techniques, advantages and disadvantages. Then will be offered own simple implementation of the recognition system.

# 1 Introduction

Speech emotion recognition has many application areas. More evident of them: robots and improving the quality of customer service.

If robot has ability to emotional speech generating, it will help him to interact with people better. Because understanding of speech emotion recognition gives understanding how to generate emotional speech. And robots with emotional speech are more people-friendly. Also, speech emotion recognition can improve the robot's decision-making system. Because there are situations when a robot needs to take into account human emotions.

Closer to reality application of this system is within an estimation of customer service quality. For example, call-centers have records of conversations. And for operator's work estimation we can identify expressive conversations and analyze them.

There are many research where described dependencies between human emotions and acoustic characteristics. Main differences between them in using different:

- Acoustics characteristics
- Features
- Classification techniques
- Recognizable emotions

The main goal of this study is to consider these differences and implement own simple solution.

# 2 Overview

## 2.1 Acoustic characteristics

Before we consider a few studies, let's consider some acoustic characteristics:

- Pitch or fundamental frequency
- Loudness or intensity
- Formants
- MFCC (Mel-frequency cepstrum coefficients)

### 2.1.1 Pitch

Human speech frequency is approximately in the range of 300 to 3400 Hz. The fundamental frequency of speech for male from 85 to 185 Hz, for female from 165 to 255.

> The fundamental frequency, often referred to simply as the fundamental, is defined as the lowest frequency of a periodic waveform. In terms of a superposition of sinusoids (e.g. Fourier series), the fundamental frequency is the lowest frequency sinusoidal in the sum [http://en.wikipedia.org/wiki/Fundamental_frequency].

### 2.1.2 Formant and MFCC

The most frequently in studies used only pitch and loudness. Although formant and MFCC characteristics also useful.

> Formant is a range of frequencies [of a complex sound] in which there is an absolute or relative maximum in the sound spectrum".[2] In speech science and phonetics, however, a formant is also sometimes used to mean an acoustic resonance[3] of the human vocal tract.[http://en.wikipedia.org/wiki/Formant]
>
> In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

## 2.2 Features

For emotion classifying from acoustic characteristics is necessary extract some features of these characteristics for every emotion. We extract and compute emotion features from training set to use it for classification. And classification is as follows:

1. Extract features from processed speech
2. Compare extracted features with emotion's features

Features can be divided to two types by how to compute and how to compare them:

- speaker-dependent
- speaker-independent

### 2.2.1 Speaker-dependent features

At first, with speaker-dependent features we need a splitted training set for every person. And compute features for every person. When classifying we need to compare extracted features with personal features. Advantages of this technique: can be more precisely and much features are not needed. These features based on some average values. For example: average pitch, average intensity, pitch range, intensity range. And they are different for every person. That is why it is not a universal technique because requires statistics keeping for each person. And it is can be applied only in a few cases.

### 2.2.2 Speaker-independent features

Unlike a previous technique here divided training set is not required. But the accuracy of classifying depends on number of speech records and number of speakers represented in the training set. Features based on speech signal dynamics, also some average values. For example, Pitch DDS (explanation further), Intensity DDS, average phrase duration, etc.