ABSTRACT
In this paper we describe a speech emotion recognition system by using k nearest neighbor classifier of statistic features of prosodic contours. We survey major approaches to emotion recognition and argue for using an algorithm dealing with a selection of statistic features of the prosodic contours with further reduction feature space by using SFFS, PCA and LDA and classification provided by k-NN classifier. We tested the designed system by using different combinations of the mentioned algorithms in order to select the optimal combination.

Categories and Subject Descriptors
I.2.7 [Artificial Intelligence]: Natural Language Processing—speech recognition and synthesis; I.5 [Pattern Recognition]: Applications

General Terms
Algorithms

Keywords
Emotion recognition, speech processing, classification, feature extraction, feature preparation.

1. INTRODUCTION
Nowadays much attention is given to speech recognition systems used in many applications such as voice recognition in navigators, voice control in mobile devices, voice search systems, etc. In most implementations the speech is converted into the text form, and then processed by using natural language processing (NLP) technologies. However it’s human to convey information not only by using words, but also with the help of emotions. There is a hypothesis that the quality of speech recognition process can be improved if the problem of recognizing emotions is taken into consideration. [16]

For emotion recognition system design Matlab environment is used because of its numerical computing orientation. For now to detect some features like pitch and formants PRAAT software is used. But in the future we plan to program all system parts on Matlab.

2. EMOTION RECOGNITION SYSTEM
An input for an emotion recognition system is a speech expected to contain emotions (emotional speech). The expected output is the classified emotion (we know that classification is the primary objective of any pattern recognition systems) [9]. The process consists of the following stages:

- Feature extraction component;
- Feature normalization;
- Feature preparation;
- Classification.

In our emotion recognition approach, we use 381 features described in table 1.

After the extraction the features are normalized by using their mean value and their standard deviation value as follows:

\[ \hat{x} = \frac{x - \mu}{\sigma} \]

Short-Term Energy Features.

Short-Term energy is one of the most important features that gives good information about the emotion [9]. It can be calculated by:

\[ E(n) = \sum_{m=n-N+1}^{n} |x(m)w(n-m)|^2 \]

where \( w(n-m) \) is the hamming window, \( n \) is the sample in analyzed window, and \( N \) is the window size.
Table 1: Extracted Features

<table>
<thead>
<tr>
<th>Indices</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 20</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range, 30, 50, 90th percentile of ST energy and first derivative of ST energy</td>
</tr>
<tr>
<td>21 - 30</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range, 30, 50, 90th percentile of ST energy</td>
</tr>
<tr>
<td>31 - 50</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range, 30, 50, 90th percentile of dB-energy</td>
</tr>
<tr>
<td>51 - 60</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range, 30, 50, 90th percentile of first derivative of dB-energy</td>
</tr>
<tr>
<td>61 - 70</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range, 30, 50, 90th percentile of rising slopes of first derivative of dB-energy</td>
</tr>
<tr>
<td>71 - 80</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range, 30, 50, 90th percentile of falling slopes of first derivative of dB-energy</td>
</tr>
<tr>
<td>79 - 85</td>
<td>Minimum, maximum, mean, range, standard deviation, interquartile range, 75 and 90th percentile of pitch</td>
</tr>
<tr>
<td>87 - 110</td>
<td>Minimum, maximum, mean, range, standard deviation, interquartile range, 75 and 90th percentile of rising slopes of pitch, falling slopes of pitch, plateaux at minima</td>
</tr>
<tr>
<td>111 - 118</td>
<td>Minimum, maximum, mean, range, standard deviation, interquartile range, 75 and 90th percentile of first derivative of pitch</td>
</tr>
<tr>
<td>119 - 142</td>
<td>Minimum, maximum, mean, range, standard deviation, interquartile range, 75 and 90th percentile of rising slopes of pitch, falling slopes of first derivative of pitch, plateaux at minima of first derivative of pitch</td>
</tr>
<tr>
<td>143 - 151</td>
<td>Jitter, RAP, PPQ5, DDP, Shimmer, APQ3, APQ5, APQ11, DDA</td>
</tr>
<tr>
<td>152 - 193</td>
<td>Maximum, minimum, mean, median, standard deviation, interquartile range, variance, skewness, 90th percentile of the 1st, 2nd and 3rd forms and BW of the 1st, 2nd and 3rd forms</td>
</tr>
<tr>
<td>194 - 201</td>
<td>Spectral energy between: 0 - 250, 0 - 600, 0 - 1000, 0 - 1500, 250 - 600, 600 - 1000, 1000 - 1500, 250 - 1000 Hz</td>
</tr>
<tr>
<td>202 - 211</td>
<td>Minimum, mean, range, median, standard deviation of voiced segments durations, unvoiced segments durations</td>
</tr>
<tr>
<td>212 - 213</td>
<td>Speaking rate, number of voiced segments</td>
</tr>
<tr>
<td>214 - 243</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range of zero-crossing rate, spectral centroid, spectral rolloff, spectral flux, spectral crest factor, spectral flatness</td>
</tr>
<tr>
<td>250 - 303</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range of 9 LPC-coefficients</td>
</tr>
<tr>
<td>304 - 381</td>
<td>Mean, minimum, maximum, standard deviation, range, interquartile range of 12 MFCC-coefficients</td>
</tr>
</tbody>
</table>

Pitch Features.

Pitch (or fundamental frequency) is connected to the possibility to distinct between male and female speeches. The pitch is determined by the frequency of vibration of the vocal cords. The pitch is individual for each person and depends on the structure of the vocal tract. There are many algorithms of pitch detection [9]: HPS, RAPT, AMDF, CPD, SIFT, etc. Pitch detection is not a trivial problem [14], so, we use PRAAT software to minimize estimation error. PRAAT is a free software for analyzing, synthesizing, and manipulating speech [3].

Formants Features.

Unlike to pitch featuring a tone of voice, Formants characterize timbre of voice. Formants are characterized by the center frequency and bandwidth [6]. Formants frequencies and bandwidth can be calculated with the help of linear prediction analysis, but in our case we use the PRAAT software in order to minimize error.

Jitter and Shimmer Features.

Jitter and shimmer are measures of the cycle-to-cycle variations of the fundamental frequency and amplitude, which have been largely used for the description of pathological voice quality [8].

In this work we use local jitter, local shimmer, relative average perturbation (RAP), period perturbation quotient (PPQ), difference of differences of periods (DDP), amplitude perturbation quotient (APQ), difference of differences of amplitudes (DDA) calculated by PRAAT. These features are described in detail in [15].

Spectral Features.

Spectral features are calculated from Fast Fourier Transform (FFT) of every short-time frame of speech signal, except spectral energy which is calculated from the whole signal. In our work we use spectral features like spectral energy, spectral centroid, spectral rolloff, spectral flux, spectral crest factor, spectral flatness which are described in [11]:

Zero-crossing Rate.

Zero-crossing rate is the weighted average number of sign changes in each signal frame. It can be estimated by:

\[ Z_n = \sum_{m=-\infty}^{\infty} 0.5 |sgn(x(m)) - sgn(x(m-1))| w(n - m) \]

where

\[ sgn(x) = \begin{cases} 
1, & x \geq 0 \\
-1, & x < 0 
\end{cases} \]

Speaking Rate.

Speaking rate is obtained by dividing number of voiced segments by number of all segments. It can be estimated by:

\[ S = \frac{N_v}{N_{total}} \]

LPC-coefficients.

Linear predictive coding (LPC) is based on the physical
model of the human speech which is presented on fig. 1. The picture shows that speech can be modelled as the output of a linear, time-varying system excited by either quasi-periodic pulses (during voiced speech), or random noise (during unvoiced speech). The basic idea of linear predictive coding (LPC) is that a speech sample can be approximated as a linear combination of past speech samples. LPC-coefficients can be determined by minimizing the sum of the squared differences (over a finite interval) between the actual speech samples and the linearly predicted ones.

MFC-coefficients.
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. So, mel-frequency cepstrum coefficients are describe an MFC.

2.2 Feature Preparation
After feature extraction we get the large set of features. If we try to classify emotions by using this set, we get very high error rate about −74%. So, let us feature set dimension. There are two known feature reduction methods:

- Selection of the most informative subset of features from the source set;
- Transformation of the source set to the new set, where result set is some function of the source set.

2.2.1 Feature Selection
We know the following hierarchy of feature selection algorithms:

- Optimal algorithms (exhaustive search, branch-and-bound, etc.);
- Suboptimal algorithms (sequential forward selection, plus-L minus-R search, sequential forward floating search, etc.).

To select the most informative set a sequential forward floating search (SFFS) was used which is suboptimal algorithm [13].

2.2.2 Feature Transformation
PCA (Principal Component Analysis) converts a set of features to a new set of linearly uncorrelated features called principal components.

LDA (Linear Discriminant Analysis) finds linear combination of features that separates object classes.

2.3 Review of Emotion Classification Methods
There are two major approaches of emotion classification [17]:

- Use of prosodic contours to classify emotions. It can be done by:
  - Artificial neural network (ANN)
  - Multichannel hidden Markov model (HMM)
  - Mixture of hidden Markov models
- Use of static features of prosodic contours. These methods are divided into two types:
  - With pdf modelling:
    - Variations of Bayes classifier
    - Parzen windows
  - Without pdf modelling:
    - k-nearest neighbor (k-NN)
    - Artificial neural network (ANN)
    - Support vector machine (SVM)

In this work we use k nearest neighbor (k − NN) algorithm for the reason that it is relatively simple and allows extending the training data set.

2.4 Database
We use Berlin emotion database [4] in order to train and test the speech emotion recognition system. This database consists of seven basic emotions: anger, boredom, disgust, fear, happiness, sadness and neutral. They are simulated using 535 speech samples.

To fit the k−NN classifier requirements we have to reduce the database: the reduced version consists of 322 emotional utterances.

2.5 Experiments
In our experiments different combinations of feature preparation algorithms were tested. For testing we use cross-validation [7]. In this method data set is divided into k (in our work k = 10) subsets and recognition procedure is repeated k times. Each time, one of the subsets is used as a testing set and other subsets are used as a training set. Recognition accuracy is computed as mean of k recognition accuracy results. Comparison of different combinations of feature preparation algorithms are shown in Table 2. It shows us, that without feature selection and preparation recognition quality is very low. Reason of that is what there are too many features, which duplicate and depend on each other. That is undesirable when k-NN classifier is used.

As you can see, even when we use the simplest feature extraction procedure, quality is improved sufficiently. The best result (accuracy of 81%) was reached with combination of SFSS + PCA + LDA which we consider a reasonably good value. Comparison with different researches results are shown in Table 3. Besides, best results are archived for anger (92%), and worst for happiness and fear (near 60%). More likely this is due to specialty of database used, emotion characteristics may also affect the result.
Table 2: Comparison of different combinations of algorithms

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>26%</td>
</tr>
<tr>
<td>SFFS</td>
<td>45%</td>
</tr>
<tr>
<td>PCA</td>
<td>63%</td>
</tr>
<tr>
<td>PCA + LDA</td>
<td>76%</td>
</tr>
<tr>
<td>SFFS + PCA</td>
<td>74%</td>
</tr>
<tr>
<td>SFFS + PCA + LDA</td>
<td>81%</td>
</tr>
</tbody>
</table>

Table 3: Comparison with other emotion recognition systems

<table>
<thead>
<tr>
<th>Research group</th>
<th>Classification</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shuller et al. [12]</td>
<td>k-NN</td>
<td>80.3%</td>
</tr>
<tr>
<td>Ayadia et al. [1]</td>
<td>HMM</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>HMM mixture</td>
<td>78.4%</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>66.5%</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>78.4%</td>
</tr>
<tr>
<td>Hendy, Farag [9]</td>
<td>PNN</td>
<td>84%</td>
</tr>
<tr>
<td></td>
<td>LVQNN</td>
<td>71%</td>
</tr>
<tr>
<td></td>
<td>BPNN</td>
<td>74%</td>
</tr>
<tr>
<td>Kotti, Paterno [10]</td>
<td>k-NN</td>
<td>75.5%</td>
</tr>
<tr>
<td></td>
<td>Gaussian SVM</td>
<td>83.6%</td>
</tr>
<tr>
<td></td>
<td>Linear SVM</td>
<td>85.6%</td>
</tr>
<tr>
<td>Busso et al. [5]</td>
<td>k-NN</td>
<td>83.5%</td>
</tr>
<tr>
<td>Our system</td>
<td>k-NN</td>
<td>81%</td>
</tr>
</tbody>
</table>

It is also worth noticing, that recognition quality depends on speaker language: best quality is achieved when both - learning and testing are executed using same language. So, when testing with polish emotion database we get accuracy just 15%.

3. CONCLUSION

In this work emotion recognition system was designed. Feature set consists of 381 features was extracted. The best result was reached with SFFS + PCA + LDA feature reduction algorithms combination and classification by k-NN classifier - 81%.

Very high recognition accuracy was reached, but it probably can be a little higher.

Despite rather good accuracy value we achieved in our experiments, there is space for further improvements in the algorithms of feature preparation and selection. We believe interesting to investigate possibilities to use support vector machines, deep learning strategy, HMM mixture [2] as a model which could improve the classification accuracy and speed.

4. REFERENCES